Chromatographic analysis of large numbers of marine bacterial extracts and the venom of the spider *Cupiennius salei*

Von der Gemeinsamen Naturwissenschaftlichen Fakultät
der Technischen Universität Carolo-Wilhelmina
zu Braunschweig
zur Erlangung des Grades einer
Doktorin der Naturwissenschaften
(Dr.rer.nat.)
genehmigte
D i s s e r t a t i o n

von Katalin Böröczky
aus Budapest, Ungarn

1. Referent: Prof. Dr. Stefan Schulz
2. Referentin: Prof. Dr. Monika Mazik
eingereicht am: 20.12.2004
mündliche Prüfung (Disputation) am: 22.02.2005
Druckjahr (2005)

Katalin Böröczky

Chromatographic analysis of large numbers of marine bacterial extracts and the venom of the spider *Cupiennius salei*

Vorveröffentlichungen der Dissertation

Teilergebnisse aus dieser Arbeit wurden mit Genehmigung der Gemeinsamen Naturwissenschaftlichen Fakultät, vertreten durch den Mentor der Arbeit, in folgenden Beiträgen vorab veröffentlicht:

Tagungsbeiträge

K. Böröczky, S. Schulz, Screening of lipophilic extracts of marine microorganisms with gas chromatography and cluster analysis, Poster, 3rd European Conference on Marine Natural products, Elmau, 2002.

K. Böröczky, S, Schulz, Handling large chromatographic data sets − analysis of lipophilic marine extracts, Poster, BioPerspectives, Wiesbaden, 2004.

CONTENTS

1 INTRODUCTION

Man has always utilized natural products for different purposes, of which medical application is the most common. Indeed, the first documents of traditional medicine trace back to the ancient ages.[1] Besides healing agents, natural products have been used as poisons, narcotics (**1**), hallucinogens, stimulants (**2**), perfumes (**3**), spices (**4**), dyes and so on.[2]

More precisely, the term 'natural products' stands for secondary metabolites; compounds produced often species specific by microorganisms, fungi, plants or animals that are not essential for the survival of the organism. The biosynthesis and degradation of secondary and primary metabolites (sugars, polysaccharides, amino acids, proteins, common fatty acids, lipids, nucleotides, RNA, and DNA) are interrelated.[2]

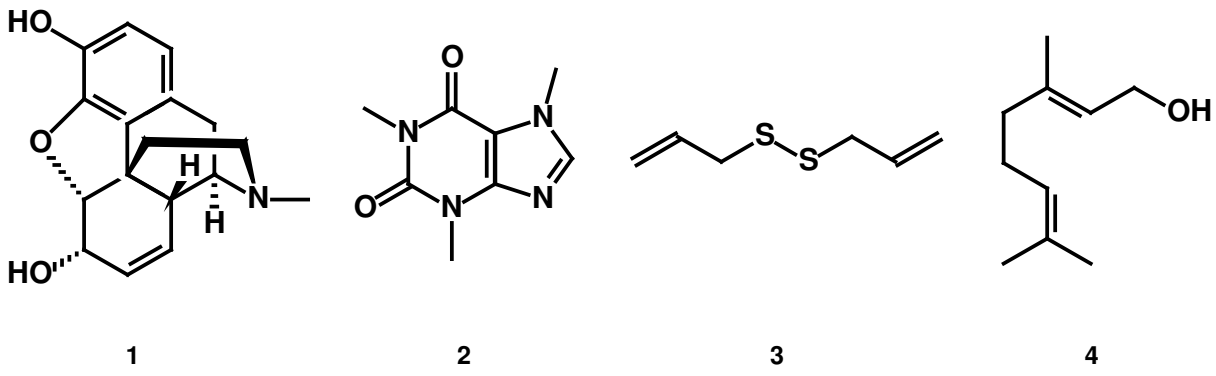
The relevance of natural product research in drug discovery is a subject of discussions in the pharmaceutical industry today. Nevertheless, according to a study in the 1990's more than 50% of the most often prescribed drugs in the US contained a natural product or a structurally related derivative.[1] Hence there is hardly any doubt about the importance of natural products.

In fact, natural product research is much more than searching for compounds that are potentially beneficial to men (as medicines, fragrances, insecticides). Who knows what kind of role antibiotics from microorganisms play in nature other than defense or which additional physiological functions pheromones may have?[3] Investigating natural products requires joint work of organic chemists, biochemists, biologists and medical scientists. It offers at the same time challenging problems to solve: ultra-trace analysis in complex matrices, synthesis of complicated structures, revealing biosynthetic mechanisms, also on the level of genetics (combinatorial biosynthesis), and exploration of substrate-receptor interactions.[3]

Early milestones of the history of natural products are records about plants and extracts used for medical purposes in the ancient world. The use of different plant oils (from *Cedrus*, *Cupressus* species, and *Papaver somniferum*) were documented in Mesopotamia as early as in about 2600 BC. Several hundred drugs, some also of animal origin were recorded in Egyptian 1000 years later, although observations began long before. From about 1100-1000 BC stem the first documents from Asia (China, India). It were the Greeks that added new facts to the field of herbal drugs in the last hundred years of the ancient western culture. This medical knowledge was used in the early centuries of the mediaeval western world. However, Arabic

and Persian scholars completed their own experience with the Greco-Roman and Asian heritage and became real experts at the same time.[1]

Personal documents about the use of natural products for healing purposes in the mediaeval ages are recipes of plant extracts from medical researchers in the 16th and 17th century. Nevertheless, the next stage of development, the isolation of active substances, such as morphine (**1**), strychnine, quinine, caffeine (**2**), nicotine, codeine, camphor, and cocaine was reported first in the 19th century.[2] It was a great challenge for chemists to work out total synthetic routes and structure elucidation techniques. These efforts lead in the 20th century to the development of research methods to explore biosynthetic pathways. Due to the results of natural product research in the last decades we know more about the ecological function of theses substances, how they contribute to the competitiveness of the parent organisms.[2]



1 **2** **3** **4**

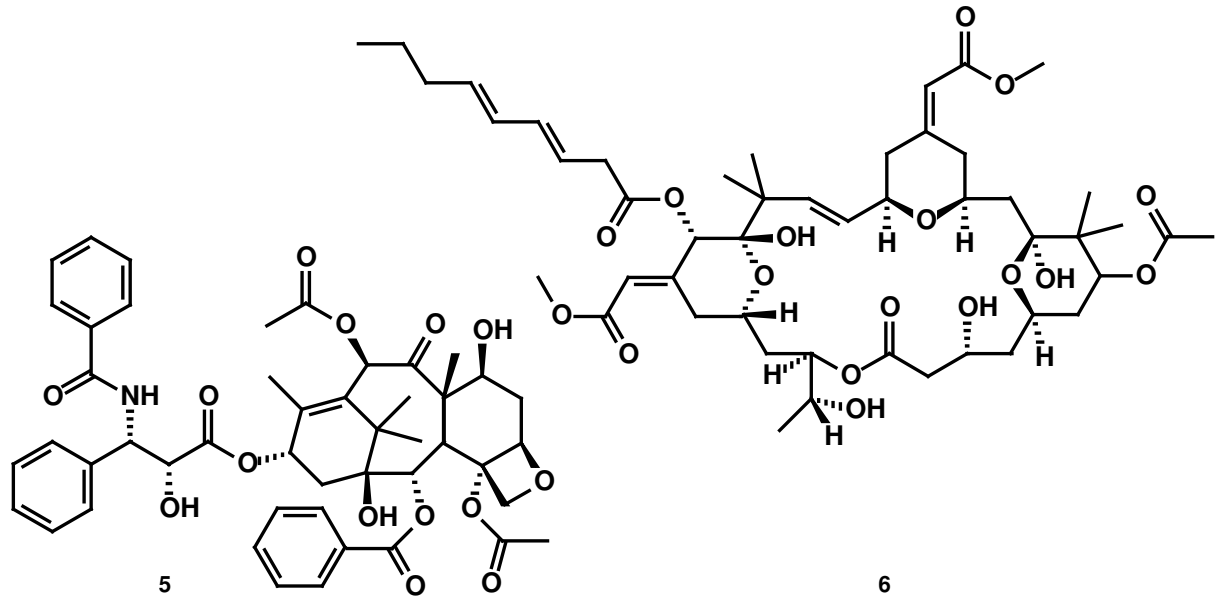
Scheme 1.1 Structures of some natural products being used traditionally: morphine (**1**) from *Papaver Somniferum*, caffeine (**2**), diallyl disulfide (**3**) in garlic, and geraniol (**4**) in rose oil.[2]

Modern pharmaceutical research based on screening of natural products began with the exploration of antibiotics. The trigger of the research work directed to antibiotics was the isolation of penicillin in 1929, followed by intense screening of bacteria and fungi after the World War II. Interest in natural product research rose again after the thalidomide scandal in the 1960's. As a consequence, secondary metabolites were applied in other pharmaceutical research fields, for instance in the development of anti-cancer drugs. One of the success stories was the discovery of the diterpene Paclitaxel (**5**), isolated from the bark of the yew tree, *Taxus brevifolia*, used under the trade name Taxol for the treatment of ovarian and breast cancer[4], and being tested against other cancer types.

In spite of the good results of natural product research combinatorial chemistry has been a major competitor in the pharmaceutical industry. Isolation and structure elucidation of the compound responsible for the desired biological activity from the natural extract found by

screening are highly time- and cost-intensive steps. In contrast, it is cheaper and faster to achieve the first clinical phase once a compound generated by combinatorial methods is proved to be active, since its structure is already known.[4] Nevertheless, combinatorial chemistry produces structures without known function in the environment.

A further problem in natural product research is the difficulty of supply of source organism or the uncertainty of future supplies, for example of marine organisms. This is especially crucial in case of complicated natural products, where synthetic production is not possible. For this reason, microorganisms that easily grow in cell culture are still preferred. Some firms (Phytopharmaceuticals, Phytera) work with plant cell cultures to avoid the problem.[4]



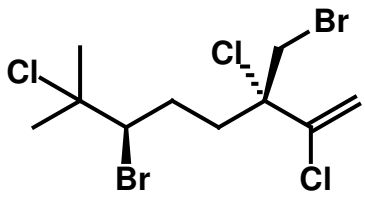
Scheme 1.2 Structure of the anti-cancer drug Paclitaxel (**5**) and the anti-leukemic marine natural product Bryostatin-1 (**6**).

As a new trend, natural product research is integrated, even though to a small percent (5-20%), in the drug discovery program of large pharmaceutical companies (Merck, Novartis). It is not the principal goal any more to find ready-to-use natural drugs, but to perform a large scale biological screening and find as many interesting substances as possible with high structural diversity. These compounds are applied as lead molecules for further investigation: combinatorial development, or chemical modification to simplify the structure or improve pharmacological properties.[1,4] In addition, smaller companies have been established to devote their efforts fully to natural product research. They usually specialize in a particular task. As a unique example, Shaman Pharmaceuticals Inc. in San Francisco organizes sample collections based solely on ethno-biological information (traditional medicine).[4]

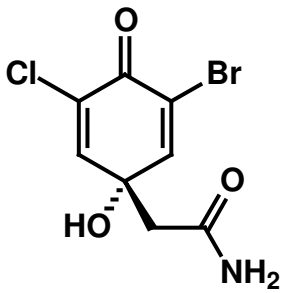
Though investments are fluctuating, the role of natural products in drug discovery remains important. It is still nature that provides the highest structural diversity.

Oceans cover 71% of the world's surface[5], and the marine biosphere has been seen lately as a newly detected source for natural products and drugs.[6,7,8] Screening for marine natural substances has been supported only since the 1970's. Development of underwater technology and ecologically relevant bioassays in the 1980's was substantial for the research. One of the first discoveries in the early 1980's was the isolation of the macrolide bryostatin-1 (**6**), an antileukemic acting substance, which is in clinical trials now. Since then, numerous secondary metabolites have been isolated and identified from all substance classes known hitherto from terrestrial organisms.[9,6]

Why should some of the marine natural products be different from terrestrial ones? The answer lies probably in the special living conditions of marine organisms. High salt concentration (30-35 g/l[9]) in an aqueous medium, absence of light, high pressure and low levels of oxygen in deep sea are the distinctive characteristics of marine environment.[10,8] As a consequence, exocrine substances are most probably water-soluble and highly active (to counteract dilution). Due to the high salinity halogenated compounds are frequently isolated from marine organisms.[8] The terpene halomon (**7**)[10] was isolated from the red alga *Portieria hornemannii*. In fact, bromophenol derivatives, the antibacterial and cytotoxic compound **8** for instance from the sponge *Aplysine cavernicola*[9], are unique to marine organisms.[11a]



7



8

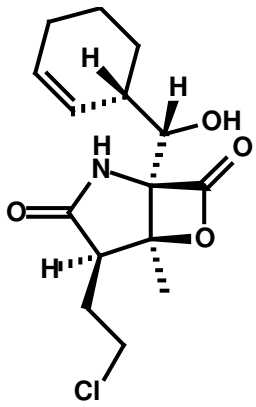
Scheme 1.3 Halogenated marine natural products: halomon (**7**) and bromochloroverongiaquinol (**8**).

The high chemical diversity of marine natural products is also a consequence of special functions of polyketide synthases (PKSs) and fatty acid synthases (FASs). These enzymes utilize catabolic products of amino acids (Val, Leu, Ile) as starter units, and propionate and butyrate as building blocks.

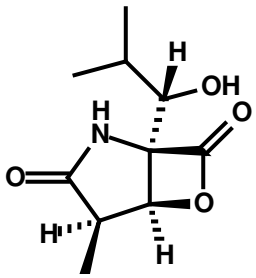
The high biological diversity in the marine biosphere is represented by the distribution of the 36 global Phyla: 34 are found in the oceans, and only 17 in terrestrial environment.[12]

In fact, around 51% of the earth's surface is covered with sea water deeper than 3000 meters;[5] and it is noteworthy that light diminishes at 200-400 m deep. Although it has been thought that deep seas merely contribute to the marine biological diversity, new diving and sampling techniques[10] made discovery of new species, the picoplankton fraction and benthic species for example, possible.

A revolutionary finding was made by the research group of one of the most devoted deep-sea scientist, William Fenical published in 2003.[13] They isolated a highly potent compound, Salinosporin A (**9**) with anti-cancer cytotoxic activity produced by a marine actinomycete strain. These bacteria species live in tropical/subtropical waters and at depths of –1100 meters. Several thousand strains of the new genus (called "*Salinospora*") isolated so far need ionic sodium for growth representing the high level of adaptation to marine environment of these species. They can be considered "real marine bacteria".



9



10

Scheme 1.4 Structure of Salinosporin A (**9**) and omuralide (**10**), a marine and a terrestrial natural product with similar biological activities.

Although the bicyclic core of Salinosporin A is identical to that of omuralide (**10**) isolated from a terrestrial *Streptomyces* strain, the functional groups are more elaborate. This is just one example for the conviction that terrestrial and marine natural products have comparable carbon backbones but the latter are often uniquely functionalized.[10] While bacteria are gaining more attention, sponges are the most frequently cited marine organisms in literature.[8] While nearly 40% of the biologically active marine compounds were isolated from sponge species in the last years, only about 2-4% was of bacterial origin excluding cyanobacteria.[6]

Nevertheless, not all bacteria are free-living, they inhabit sediments and the surface or internal spaces of other organisms, sponges for example. The relationship to the host organism can be symbiotic or pathogenic.[9] Indeed, often the symbiotic microorganism is the producer of natural products isolated from invertebrates.[6]

The development of marine natural product research is probably slower than expected due to particular difficulties. It takes more effort and is more expensive to collect samples. There are no examples among traditional medicine records for the use of substances of marine origin to help to find potent organisms. The concentration of interesting substances is often very low, which makes isolation and structure elucidation more complicated.[10]

Another barrier is that culturing marine bacteria is troublesome. Due to the "great plate count anomaly", i.e. only a tiny part of the microscopic organisms build colonies on plates,[14] the estimated fraction of already studied microorganisms is less than 0.1%[15]. Considering that the number of microorganisms is about 10^5-10^6/ml in sea water, there is a lot to do.

The chemical ecology of spiders is a significantly narrower research field compared to studying marine natural products; and it has been by now thoroughly investigated. However, owing to new analytical techniques, for instance high performance liquid chromatography coupled with diode array detector and tandem mass spectrometry (HPLC-DAD-MS^n) further substances can be identified from already known species.[16,17] In addition to more sensitive analytical methods, novel milking techniques providing larger quantities and better quality of spider venom contribute to this trend. Indicating 'the steady-state' of the research only a small number of pharmacological firms are interested in terrestrial invertebrates.[4]

Although the majority of the substances will find probably no practical application in human therapy,[11b] screening natural products for cytotoxicity are the most granted research programs.[8] This may distract attention from the ecological role of natural products other than defense.

Secondary metabolites have diverse functions in nature, which are still to be discovered. This is especially true for marine natural products and marine bacteria. More specifically, the investigation of lipophilic compounds of marine origin is rather negligible compared to other substance classes.[6] An interesting environment to investigate is the deep sea, since non-costal regions are probably free of pollution[9,5] and the influence of terrestrial life[11].

Exploration of the high complexity of natural products requires constant technical development, which again, results in discovery of novel structures. This automatism is the key of the future success of natural product research.

2 OBJECTIVE

The aim of this Ph.D. work was to develop an analytical tool for the investigation of several 100 marine extracts obtained from Prof. Laatsch, from the University of Göttingen, Germany, and to examine several low molecular weight (LMW) fractions of spider venom by capillary electrophoresis (CE) in cooperation with Prof. Kuhn-Nentwig from the University of Bern, Switzerland.

Focus of the first part of the work was on the high-throughput analysis of gas chromatographic (GC) data of lipophilic marine fractions, usually discarded by chemists. A conversion program was to be written in order to enable the use of the statistical program package SPSS for the clustering of the samples based on their GC patterns. Different modes of cluster analysis were to be examined concerning hierarchical cluster analysis, principal component analysis, as well as various distance measures and linkage rules. Isolation and structure elucidation of novel substances found by the new method and the presentation of other interesting compounds should prove the successfulness of the whole process. The influence of fermentation conditions was also to be studied. The lipophilic fermentation extracts were investigated within a research program in Lower-Saxony, Germany.

Crucial point in the second part of the work was the application of CE for the analysis of some LMW fractions of the venom of *Cupiennius salei*. Samples were to be examined first by derivatization and gas chromatography mass spectrometry (GC-MS). Furthermore, the usefulness of CE and indirect UV detection for the analysis of LMW venom components was to be investigated. Results were to be compared to literature data concerning spider venom in general.

3 MARINE BACTERIA

3.1 Research project

Introduction

The analysis of lipophilic extracts from marine bacteria was started within the Marine Biotechnology Research Project of Lower-Saxony, 1998-2002. The aim of the project was to investigate bacterial natural products in the North Sea and the Wadden Sea, in Germany. Several groups from different universities and institutes from Lower-Saxony worked together in this basic research program. The combined efforts of scientists from the field of biology, molecular biology, natural product research, organic chemistry, pharmacology and biotechnology focused on the identification of biologically active substances.

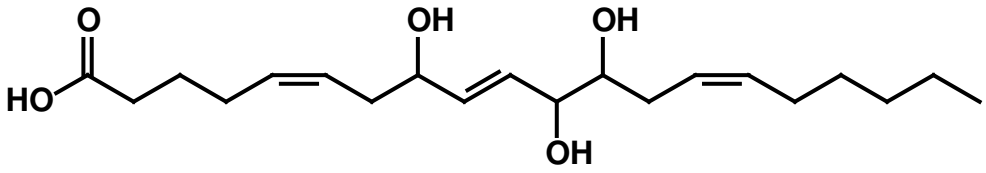
The research activities were divided into five big subject areas. In the first one, chitosan and chitosan oligomers (I), enzymatic pathways were investigated for the environment friendly production of chitosan. The second and third subject areas, competition on surfaces (II) and secondary metabolites of microorganisms from the North Sea (III), were interrelated to a high degree. Several macroorganisms (tunicates, bryozoa) and the associated bacteria were isolated and characterized according to phylogenetical and physiological properties. Biological activity of the secondary metabolites were tested. Interesting compounds were synthesized for testing purposes by the research groups in the forth subject area, synthesis of marine natural products (IV). The aim of the fifth subject area (V) was the biological testing of samples and the development of new test methods. Antibiotic, anticancer and antiviral activities of newly found substances were analyzed and a voltage-dependent ion-channel test system was established.

Contribution of our research group to the project in the first time period 1998-2000

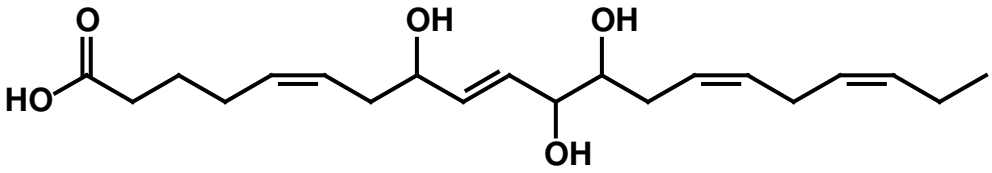
Our group participated in the work of two subject areas of the project: Synthesis of Marine Natural Products (IV) and Secondary Metabolites of Lately Discovered Microorganisms from the North Sea (III).

Dr. Markus Müller was involved in subject area IV with his work on the egg-hatching-factor of barnacles.[18] The compounds **11** and **12**, also called trioxilin A_3 and A_4, are the two active

constituents of the egg-hatching-factor of the barnacle, *Balanus balanoides*. (Scheme 3.1) However, the stereochemistry of the hydroxyl groups had not been elucidated. Trioxilin A_3 and A_4 were found in the extract of the test organism, *Elminius modestus*. Saturated derivatives of the acids with all the four pairs of diastereoisomers were synthesized in order to find out which stereoisomer is produced in nature.



11

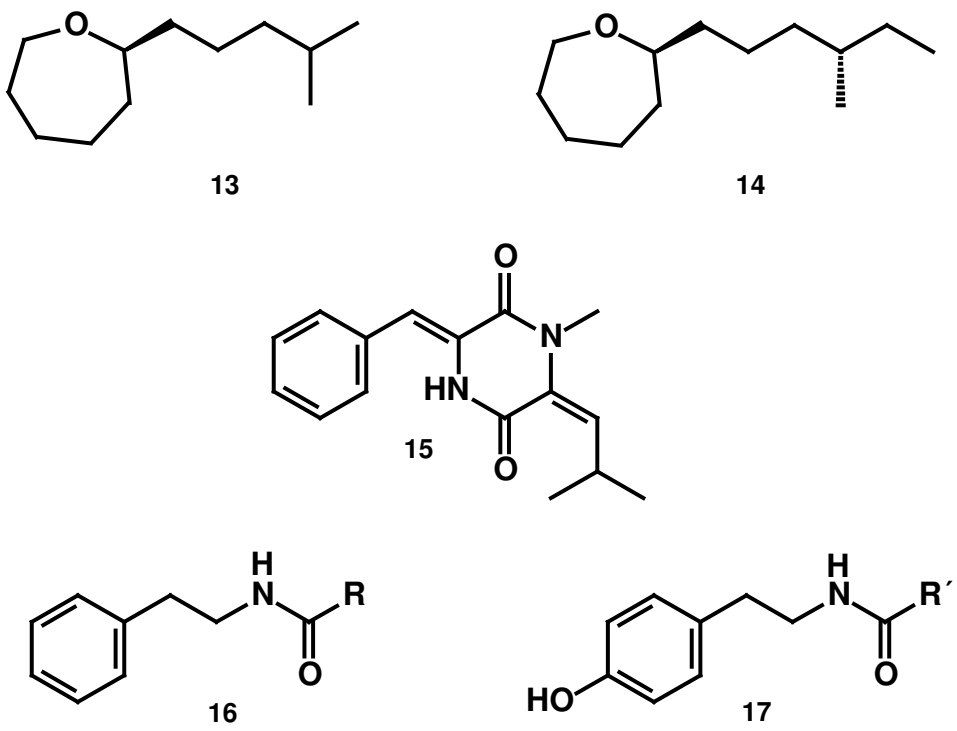


12

Scheme 3.1 Trioxilin A_3 (**11**) and Trioxilin A_4 (**12**).

The four diastereomers and seven of the eight enantiomers were separated on achiral and chiral GC phases. The natural product contains three or four isomers with an *anti*-(11,12) stereochemistry and the main isomer has a (8*R*,11*RS*,12*RS*) configuration.

Another task was to investigate the lipophilic fraction of the fermentation extracts of marine bacteria by collaborating with groups from subject areas II and III. In the first period extracts of marine *Streptomyces* strains were analyzed after derivatization by GC-MS by Dr. Katja Stritzke.[19] Beside ω-1 and ω-2 branched fatty acids, a new natural product, 3-hydroxy-12-methyltetradecanoic acid was found in a sample from a strain. Other constituents, mostly already known from S*treptomyces* strains, were aromatic acids (benzoic and antranilic acid), hydrocarbons and diketopiperazines, for example. Further compounds were identified by synthesis and structure elucidation, summarized in Scheme 3.2.



R = **a** methyl, **b** ethyl, **c** isopropyl, **d** propyl, **e** 2-butyl, **f** isobutyl, **g** butyl, **h** isopentyl, **i** pentyl, **j** 3-methylpentyl, **k** 4-methylpentyl

R´ = **a** methyl, **b** ethyl, **c** 2-butyl, **d** isobutyl, **e** isopentyl

Scheme 3.2 Natural products identified from marine strains of *Streptomyces*.

The caprolactones **13** and **14** have a growth inhibitory effect on tumor cells.[20] The compound **x** is a known stereo isomer of Albonoursin, and has antibacterial and cytotoxic activities.[21] Furthermore, different *N*-phenylethyl-alkylamides, **16,** and *N*-substituted tyramine derivatives, **17,** were identified by GC-MS. The tyramine **17d**, for example, is known from a corynebacterium and has an inhibitory effect on aldose reductase.[22]

The second time period of the project, 2000-2003

The investigation of lipophilic marine extracts proceeded further in the second period of the project. After preliminary experiments several hundred marine strains were isolated and fermented. Our group received lipophilic fractions of fermentation mixtures produced mostly in the research group of Dr. Laatsch in Göttingen. Isolation of the strains was carried out by the group of Dr. I. Wagner-Döbler from a biotechnological institute, 'Gesellschaft für Biotechnologische Forschung', in Braunschweig.

Marine strains were isolated from the North Sea in different ways.

1. Seawater was taken 10 m deep near Helgoland and single colonies were cultivated after a series of dilution steps (samples Hel1 to Hel85).
2. 100 liters of water sample were filtered (5 μm) and the filtrate was kept in the dark at 4 ^{o}C for enrichment of bacteria. The picoplancton fraction, obtained by a further filtration step (0.2 μm), was also investigated (samples Picx).
3. Sterile plates of glass were kept 0.5 m deep in the seawater, 1 km from Helgoland; the biofilm was scratched off and the bacteria were isolated.

Isolated bacteria were first cultivated, than fermented to study the production of secondary metabolites. It is extremely difficult to cultivate marine bacteria. As a result, several types of medium had to be investigated and used throughout the whole project. Due to the diversity of medium constituents (sugars, proteins, dextrin, fish-meal, algae extract, metals, vitamins, calcium, for example) the extracts contained a lot of compounds that led to a high background in the chemical screening. The extraction yields were also rather low at the beginning, which resulted in very low amount of sample available for chemical analysis.

Fermentation products were usually separated between methanol and cyclohexane, and the latter fraction was investigated in our laboratory by GC and GC-MS. Selected strains and control samples without bacteria were fermented in parallel experiments. These fermentation mixtures were extracted with ethyl acetate, only.

In addition to marine samples, extracts of terrestrial *Streptomyces* strains (with a label GWx/x) were also obtained from Mr. Laatsch, from Göttingen.

Due to the high number of extracts to analyze a selection tool had to be developed using data analysis based on clustering of the GC patterns.

3.2 Multivariate data analysis

In the following chapter I am going to discuss the theoretical background of data analysis of samples based on their chromatographic properties.

Handling large data sets has become routine work for chemists as for other scientists. The number of the samples in a study is often a few hundred, or even more. The dimension of measurement data has also increased over the years with the development of chromatographic and spectrometric techniques. To find relationship between multiple measurement data and not directly measurable sample properties is impossible without computer assisted multivariate analysis.

Data mining and pattern recognition procedures use a collection of multivariate data analysis methods. A pattern is a set of features that describe certain properties of an object. Pattern recognition means identifying a pattern as a member of the class to which it belongs.[23] Samples are usually the objects and measurements the features in chemical applications (section 3.2.2).

Two approaches for data analysis are unsupervised and supervised learning. The first is distinguished from the second by the fact that there is no outcome measure, the goal is to describe the associations and patterns among a set of input measures. Supervised learning is a technique for predicting the value of an outcome measure based on a number of input measures.[24]

As a rule, each sample of a raw data set is considered as a point in an n-dimensional measurement space. The samples, $i = 1$ to m, are represented by the data vectors, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ij}, \ldots, x_{in})$, where x_{ij} is the j-th measurement of the i-th sample, the area of the j-th peak in the i-th chromatogram, for example. The distance of two objects in the measurement space can be calculated (section 3.2.1.1). The smaller the distance, the more similar the samples. The method of choice for data analysis depends on what we are interested in: the natural grouping of samples (cluster analysis) or reduction of dimensionality (principal component analysis).

3.2.1 Unsupervised data analysis

Since we had no *a priori* knowledge about our data set unsupervised analysis techniques were applied to extract information. The major methods: cluster analysis (CA) and principal component analysis (PCA) are presented below.

3.2.1.1 Clustering

Cluster analysis is a collection of methods to organize data by uncovering the natural grouping of samples or variables in a data set. Similar objects are grouped together and thus form a cluster. Most widely used is the so called hierarchical cluster analysis (HCA) of samples, particularly the agglomerative form of it. At the beginning samples are considered as one-member-clusters and the distance between each pair is calculated. Then the two nearest clusters are merged into a new cluster. The distances are calculated again with the new cluster, and the clustering proceeds, till all of the samples are assigned to a cluster.

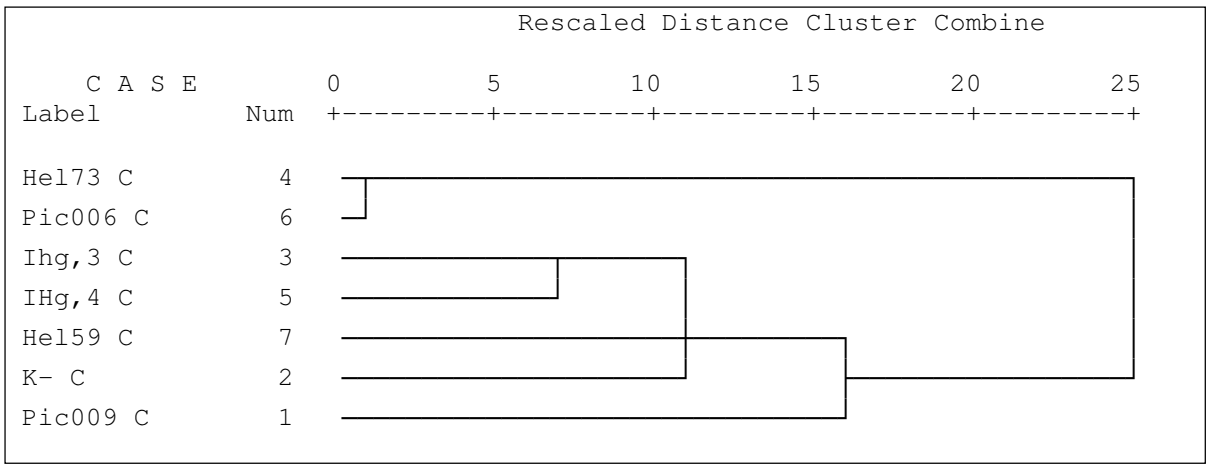


Fig. 3.1 A dendogram made by Euclidean distance and average linkage between groups. Hel73 C and Pic006 C are the two most similar objects.

The clustering is visualized by a tree-like diagram, called dendogram. The position of the junction points are representing the level of similarity. In Fig. 3.1, for example, the samples Hel73 C and Pic006 C are the most similar ones. The clustering process is time intensive and is suitable for a few hundred samples only.

Other clustering methods are two-way joining and K-means clustering. The first one is useful for discovering grouping within objects and variables simultaneously. The second is for organizing objects into a given number of clusters.

Distance measures

The metric used for the calculation of the distance between two points in the measurement space, called distance measure, depends on the data type. Binary data have only two values (1 or 0, or yes or no), discrete data are represented by a finite number of values (taste of apple: very good, good, moderate, unsatisfactory). Chemical data are mostly continuous, and the number of values is infinite. Data scale is also important; the area of chromatographic peaks, for example, is a ratio scale continuous data type, since the scale has an absolute zero and ratios are meaningful. Possible distance measures for this type of data are:[25a]

$$d = \sqrt{\sum_{i=1}^{p} (x_i - y_i)^2}$$ Euclidean distance,

$$d = \sum_{i=1}^{p} (x_i - y_i)^2$$ Squared Euclidean distance,

$$d = \sqrt{\sum_{i=1}^{p} |x_i - y_i|}$$ City-block (Manhattan) distance,

$$d = \max |x_i - y_i|$$ Chebychev distance, and

$$d = \left\{ \sum_{i=1}^{p} (x_i - y_i)^m \right\}^{\frac{1}{m}}$$ Minkowski distance,

where x_i and y_i are the i-th coordinates of samples x and y. The Euclidean distance provides the smallest distance between two points in an n-dimensional space. It is suitable for raw data, nevertheless, variables differently scaled are to avoid.[26] Larger distances are weighted by choosing the squared Euclidean metric. City-block distance is more suitable for discrete distances. In order to save computational time Chebychev distance is used. Minkowski is identical to city-block at $p = 1$, Euclidean at $p = 2$, and approaches Chebychev as p increases.[25a]

Linkage rules

The way the distance between clusters (with more than one members) computed, called linkage method, has an effect on the final structure of the clustering. The different linkage methods are presented in Fig. 3.2. For single, complete, and average linkage the distances between all pairs of points are calculated, taking points from different clusters.

By single linkage the smallest distance, by complete linkage the largest distance is considered as distance between two clusters. The first one results in chain-like clusters and is sensitive to outliers, the second one is more suitable for distinct, compact clusters.[24b,25b]

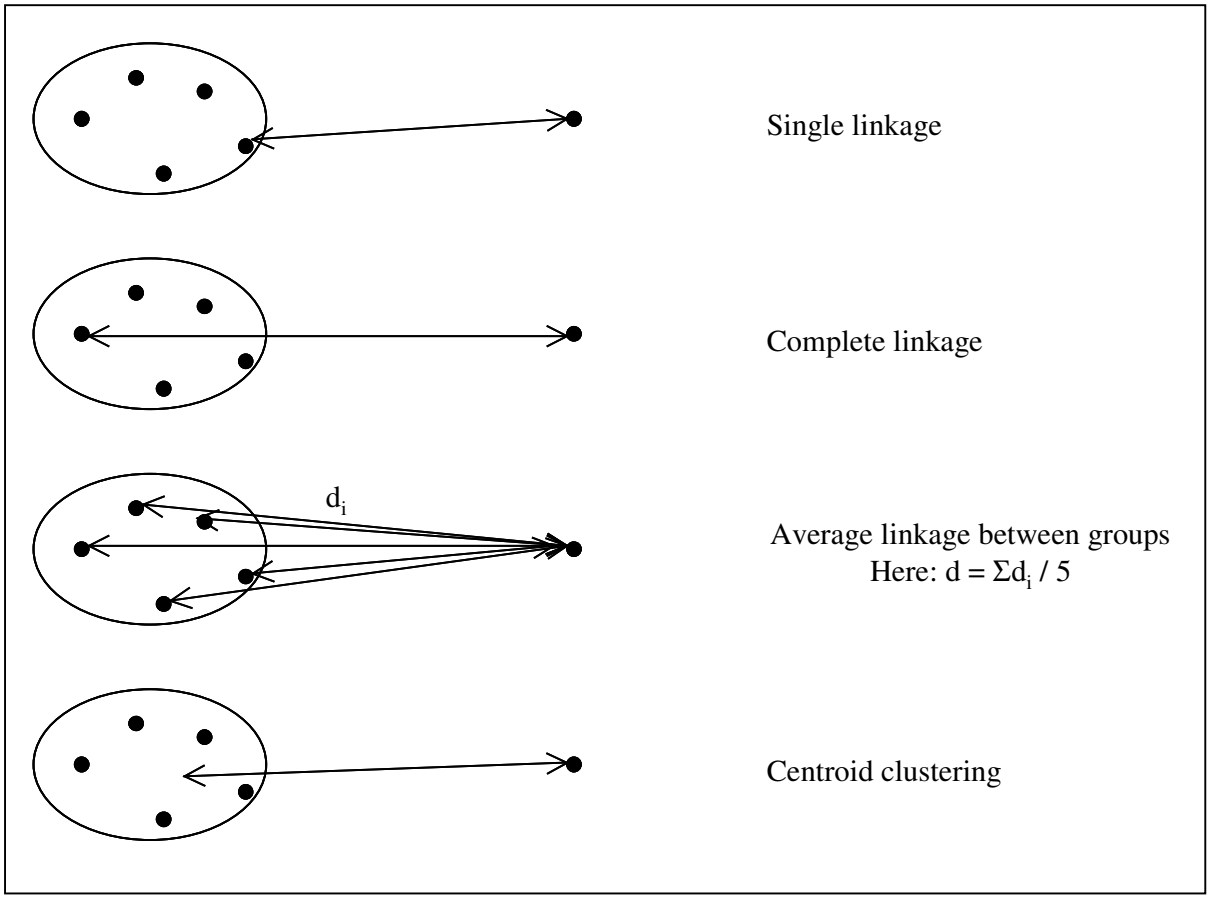


Fig. 3.2 Linkage methods (for the Ward method see text).

The average linkage joins clusters according to the arithmetical average of the distances and is generally good in most cases.[24b,25b] Weighted average linkage fits better, when the size of the clusters varies in a large scale. The centroid is a point in the measurement space with coordinates that are the weighted or unweighted arithmetic average of the coordinates of all points in the cluster. For centroid methods the centroid of the clusters is calculated; the

distance between two clusters is the distance between the centroids. The Ward method is a different one (not presented in Fig. 3.2); those clusters are fused that cause the smallest increase in the sum of the squared Euclidean distances of the points to the centroid of the cluster. It tends to produce small clusters.[26] Centroid and Ward clustering methods are used together with the squared Euclidean distance.

Interpretation of the results of the hierarchical cluster analysis

The limit of cluster analysis is the definition of the threshold value of similarity.[27] It depends on the aim of the experiment, how many clusters we recognize on the dendogram. It is difficult to tell how well the outcome of the hierarchical clustering reflects the real grouping in the data set. In Fig. 3.1, for example, only two clusters are seen at the distance level 17, but five at the distance level 8.

An indication for well separated clusters in a data set is that each linkage method gives the same distribution of clusters. In case of overlapping clusters the clustering should be performed with several linkage methods simultaneously.

3.2.1.2 Principal component analysis

In cases when variables are correlated to some extent in a raw data set it makes sense to find a smaller number of independent variables that contain almost all the information than the original ones. As a result, grouping of the objects or the original variables can be detected. Nevertheless, the main purpose is to reduce the dimensionality of the data. The collection of these methods is called data reduction.

The most widely used method is the principle component analysis. The original n-dimensional measurement space is transformed to another n-dimensional one, where the first few dimensions (components) hold most of the information (variance). The components are calculated by finding the eigenvectors (and eigenvalues) of the correlation matrix of the original variables. The first component represents the highest variance in the data and has the largest eigenvalue. The second one has the second largest eigenvalue, and so on. Ideally the first few components, the so called principal components, contain a high percentage of the total variance and the rest of the components contain noise. Thus the principal components can be extracted without major loss of information and they form the new axes in the measurement space. The coordinates of the objects are called scores, and that of the variables

loadings. Rotation of the axes gives the final solution, where the correlation of the components is minimized.

The number of components to extract can be determined differently. A minimum eigenvalue the components should have can be fixed. The number of components to extract can also be selected before the process, nevertheless, it is better to perform preliminary analysis to estimate it (with the first criteria, for example). It is also useful to draw a so called scree diagram, the eigenvalue versus component number (Fig. 3.3). The elbow of the plot shows the number of components that are necessary to describe the data set. A more exact criterion is the minimum portion of the total variance the principal components should represent. Extracting 75 % of the variance leads usually to satisfactory results. The parallel use of the last two mentioned methods is advisable.[28]

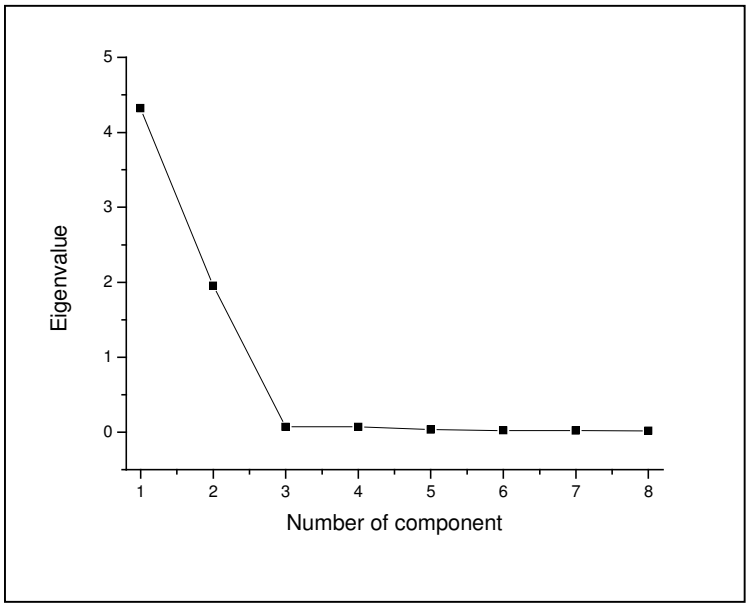


Fig. 3.3 A scree diagram. The estimated number of principle components in this case is two.

Interpretation of the results requires caution. Grouping of the variables can be visualized by drawing 2D score plots with selected pairs of principal components as the two axes. Then, the meaning of the grouping of the variables is examined. It is not always possible to assign a property to each group. To discover the grouping of objects with PCA is more productive, if some structure in the data set is already known. Classification methods involve PCA steps to reduce dimensionality and both CA and PCA for discovering data structure, as presented in the next section.

3.2.2 Applications in chemistry

'Computer pattern recognition methods extend the ability of human pattern recognition, but, in the end, it is the chemist who must do the chemistry.' wrote Kowalski in 1975.[29]

He was one of the pioneers who applied pattern recognition for chemical purposes. The development of these methods was very fast in the '70s owing to the big interest in predicting molecular structure from spectral (mainly MS) data. Since 1974 Chemometrics has been the name of this special field of chemistry.[23]

The application of clustering and/or PCA for the classification of chromatographic data is remarkably widespread: for quality control purposes in the pharmaceutical and biotechnological industry, for diagnosis in medicine, for sensory evaluation and quality control in the food industry, for identifying oil source in the petrol industry, for environmental studies, and to determine the origin of objects according to some chemical properties, just to mention a few.[30]

A few applications are presented to highlight some factors that influence the successfulness of classification of chromatographic data and to demonstrate the development of the methods.

The way chromatographic peaks are integrated can affect the results, for example. Nonenzymatic posttranslational modification of collagen was followed by capillary zone electrophoresis in a study.[31] Samples from groups of differently treated genetically hypertriglyceridemic animals and a control group were investigated. Due to the lack of baseline separation and the poor reproducibility of migration times not the single peaks but segments of the electropherogram served as variables. Four of the seven segments were enough to discriminate the different groups of animals. The results of the PCA was dependent on the peak integration method: valley-to-valley or common baseline. The second method gave no meaningful results.

The standardization of raw data and the selection of the linkage method used for hierarchical cluster analysis is also relevant. In a research work from 1984 urine samples from patients with different renal disorders were clustered according to the HPLC analysis of the urinary proteins.[32] Hierarchical cluster analysis using single linkage was not successful owing to the overlapping of the clusters. The dendograms obtained by Ward's method, once with standardized data, once with the raw data, were different. Nevertheless, the interpretation was possible and the grouping was clinically meaningful in both cases.

The use of CA and PCA to reveal taxonomic relationships from chemical data should be taken with caution. An Australian research group examined the gas chromatograms of some

essential leaf oils gained by a vacuum distillation technique.[33] Although the three *Eucalyptus* series are clearly differentiated using morphological observations, no meaningful grouping was achieved by CA or PCA based on chromatographic data. The variation in the leaf oil composition of species was so large between and within series that the three series could be differentiated only by supervised learning using linear discriminant analysis.

A recent publication describe the classification of samples using 2D measurement data. Jet fuels were investigated by a 2D gas chromatographic technique in the United States.[34] The 2D separation, called comprehensive GC x GC, is achieved by transferring little gas plugs from the first column onto the second one. PCA and K-means clustering was performed on the data set after selecting the relevant features. Not only the jet fuel types were successfully differentiated (independent from geographical origin), but also the fuel mixtures with different composition.

Kowalski was right, interpretation of the results is impossible without human sense. However, there exists the danger of overfitting parameters to obtain results that fit better the expectations.

3.2.3 Introduction to SPSS

SPSS Inc. exists since 1975 as a company, but its history started more than 35 years ago.[35] Three young men, a social scientist, an expert in operational research and a business administrator developed a software system to quickly analyze volumes of social science data, in 1968. It was then called Statistical Package for the Social Sciences. Having their office at the University of Chicago, they worked continuously on the development of the software and sold it mostly to universities. After incorporating in 1975, the circle of the customers became larger. The first statistical package for PCs appeared on the market in the mid-1980s. In 1992, SPSS Inc. provided the first statistical product for Microsoft Windows systems.

The term SPSS covers Statistical Product and Service Solutions now. The company has a large market penetration with its data mining products and concentrates on predictive analytic applications in the future.

The software SPSS is an all-round statistical tool for different sciences.[36] It offers data transformation, descriptive and explorative data analysis, data tests, variance analysis, correlation and distance calculation, linear regression analysis, cluster analysis, discriminant analysis, factor analysis and multidimensional scaling, for example.

Input data can be edited or undertaken from other programs. SPSS accepts Excel worksheets and also data in ASCII format. After selecting a procedure from the menu, the variables are chosen and the parameters are set by the user. The results are presented on a new sheet, which can be saved as a separate file. It is also possible to proceed further with the analysis of the results.

3.3 Development of a high throughput process for the gas chromatographic investigation of several hundred marine extracts

It was our task to explore the lipophilic compound diversity of marine bacteria from the North Sea and the Wadden Sea within the marine research project in Lower-Saxony (section 3.1). Hundreds of samples containing hundreds of components with concentrations from trace to high level had to be analyzed. The chemical screening of the samples required a high efficiency separation method coupled with universal detection for the identification of the components. The method of choice to separate and identify substances in marine extracts was gas chromatography-mass spectrometry. A large scale of polarity of the analytes was to be expected, partly due to the constituents of the cultivation media. Hence, it was necessary to derivatize the extracts. Since the number of samples was too high to analyze all of them by GC-MS, a data selection tool was developed based on GC data. The tool implies a self-made program, ChromConv to convert GC data and a clustering step.

The combination of GC and GC-MS has advantages over HPLC-MS for the fast chemical screening of hundreds of unknown samples for LMW compounds, under 1000 amu. (Table 3.1). The development of a separation method is easier, since a single BPX-5 capillary provides the separation of unpolar to mid-polar substances with high theoretical plates. Only a temperature program has to be created before analysis, instead of searching for the proper mobile phase mixture and gradient. The electron impact (EI) ionization in GC-MS is universal for analytes with a molecular weight below 1000 amu, whereas it is time consuming to find the optimal ionization parameters for the typical HPLC interfaces (ESI and APCI). Successful trace component analysis by GC-MS is due to the good overall sensitivity and the possibility of single ion monitoring (SIM). Tandem HPLC-MS techniques make analysis in trace level possible, but again, the method development is time consuming. In addition, the problem of

analyzing more polar substances by GC-MS can be solved with a proper derivatization method (section 3.3.2.1). Considering the small amount of extracts obtained in the project, often less than 1 mg, GC-MS remained the technique of choice for the chemical screening.

Table 3.1 Comparison of GC-MS and HPLC-MS techniques for the analysis of unknown mixtures containing low molecular weight compounds (<1000 amu).

	GC-MS	HPLC-MS
Costs	inexpensive	expensive
Analytes	volatile, thermally stable, but derivatization helps!	nearly without limitation, but no universal ionization
Speed of separation	fast	slow to fast
Theoretical plates	10^5-10^6	10^4-10^5
Resolution	unique	good
Enhanced sensitivity by	selected ion monitoring	tandem MS
Method development	fast	slow

Since the number of samples to analyze reached quickly several hundred, the complete characterization of all of them by GC-MS was unthinkable. Traditional GC-MS analysis requires up to 5 different runs including different derivatization experiments. Still, we wanted to get an overall picture about the lipophilic compounds produced by marine bacteria. Moreover, it was our intention to isolate and characterize novel marine substances from selected strains.

Therefore a data handling tool was developed to find grouping in our data set based merely on the results of gas chromatographic measurements. In order to be able to compare the results, all extracts were trimethylsilylated under standardized conditions before GC analysis. Without any knowledge about the data structure unsupervised data analysis was performed on the GC data. A GC analysis is inexpensive and fast. It was possible to derivatize and measure all the samples by GC. Raw GC data was converted by the self-made program, ChromConv. Clustering of the samples based on the converted GC data was performed by HCA and PCA/HCA, thus grouping of similar samples was revealed.

For the identification of substances by GC-MS samples were selected with the help of our data selection tool. It was then possible to avoid the GC-MS analysis of very similar samples (dereplication). The other aspect was selecting interesting extracts that probably contain new metabolites.

3.3.1 The program ChromConv

ChromConv is a data conversion program. It takes raw GC data as input and creates a so called measurement matrix in form of a text file. The output file contains a table with samples in rows and variables in columns and can be used for clustering.
Perl was the language of choice for writing ChromConv.

3.3.1.1 Compensation for retention time changes

It was important to take into consideration how the retention time of a given compound changes during the years of the experiment. Retention time fluctuation was followed by regular analysis of a standard mixture of alkanes from $C_{11}H_{24}$ to $C_{44}H_{90}$.

$$RT_{i,corr} = RT_i + \Delta RT_i \qquad \text{Eq 3.1}$$

The retention time deviation from the first value measured at the beginning of the experiment was plotted against the actual retention time. Polynomial fit of the second order was applied to the points. (Fig. 3.4) As a result, the retention time values were corrected in the whole measurement time interval according to the parameters of the fitting curve (see Eq. 3.1).

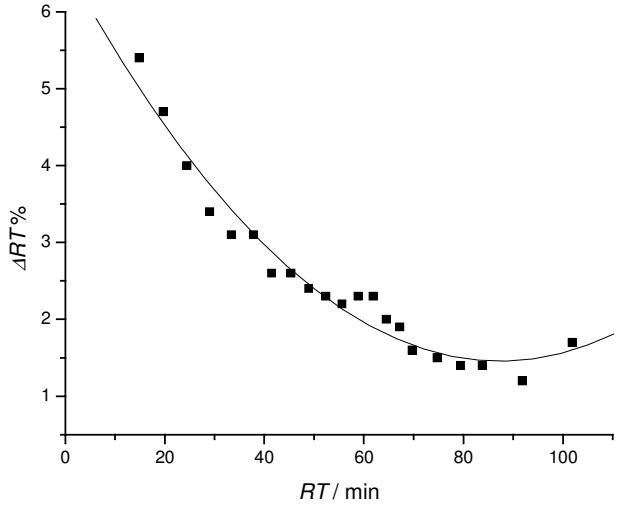


Fig. 3.4 Retention time changes of the standard alkanes versus retention time (compared to the values at the beginning of the experiment).

3.3.1.2 Conversion of gas chromatographic data

The output of the conversion program had to be a two dimensional matrix with samples in rows and variables in columns. It was quite obvious that each marine extract or rather the GC chromatogram of it represents a case. The question was, which chromatographic feature holds the information for the classification of the samples.

Since we had unknown samples and the GC analysis did not provide identification it was obviously impossible to compare samples according to the existence of given substances. To find the feature for the differentiation of the chromatograms, the format of the measurement results was examined. GC data are stored in two different ways by the measurement software (ChromCard) supplied from the manufacturer of the GC system. (Fig. 3.5) One form is a list of intensity values and the points in time it was measured, the raw GC data. The points in time are the same for each chromatogram by a given data rate. The other form is the measurement report. It contains the retention time and the integrated area of the peaks listed in pairs.

Data alignment, that is peak matching with a given deviation or time window, is quite common.[32,37,38] However, significant shifts of retention time, nearly 0.5 min, could be observed due to the lack of any carrier flow controlling system in the GC. The shift was not constant in the measurement time interval due to the splitless injection mode used in the experiment. In addition, the deviation was comparable to the retention time difference of two adjacent peaks.

Another approach is the conversion of raw GC data.[37] Clustering of samples occurs then according to the intensity value measured at a certain point in time. Disadvantage of this method is the loss of chemical information, since the chromatograms are considered as graphic lines.

Finally, a modification of the first mentioned strategy, retention time-peak area concept, was used. The measurement time axis of the chromatograms, 6-110 min, was partitioned into equally large time segments. This way the dimensionality of the data is controlled.[39]

The length of a segment was typically 0.5 or 1 minute. It was large enough to compensate the high retention time deviation (see text above), but not too large to lose too much from chemical information. The integrated peak area values were summed up within each time segment (Eq 3.2a and 3.2b) and normalized optionally to the maximum segment value (Eq. 3.3) or to the sum of all peak area values (Eq. 3.3b) for each chromatogram. Peaks in the first section of the chromatogram (6 minutes) were ignored, since only the solvent and some derivatization by-products eluted so early from the column.

Ascii chromatogram data

Time(Min)	Datapoint(µV)
.	
.	
.	
47.75	1722.
47.7667	1723.
47.7833	1724.
47.8	1733.
47.8167	1752.
47.8333	1764.
47.85	1772.
47.8667	1781.
47.8833	1817.
47.9	1904.
47.9167	2063.
47.9333	2365.
47.95	2830.
47.9667	**3220.**
47.9833	3167.
48.	2724.
48.0167	2207.
48.0333	1886.
48.05	1758.
48.0667	1719.
48.0833	1708.
48.1	1707.
48.1167	1703.
48.1333	1708.
48.15	1709.
.	
.	
.	

Chrom-Card Report

Operator ID : User 7 Company Name : AK Schulz
Method Name : Katalin D Method File : KATADNAT.MTH
Analysed : 06-06-01 22:01 Printed : 05.09.2001 11:27
GC Method : 0 Sampler Method : 1 Vial # 3

Sample ID : 1/IHg,4 (# 102) Channel : Channel D
Analysis Type : UnkNown (Area) Calc. Method : Area %
Chromatogram: C:\BENUTZER\KATALIN\DATA\kbnat102.DAT

Retention Time (Min)	Area (æV/Sec*10)
47.97	**62969**
51.35	207724
67.18	31529
67.60	105240
67.84	131490
68.07	184831
68.30	234201
68.46	772874
68.86	29095040
68.89	4155142
73.51	377798
77.98	122723
79.02	66852
83.16	2141683
83.26	88274
87.17	4419515
90.87	2999515
94.57	726995
	45924400

Fig. 3.5 Two GC data formats available by ChromCard. To the left: part of a list of raw GC data from a measurement; data rate is 60 / min. The peak is marked at 47.9667 min. To the right: measurement report of the same sample; retention time and area of the same peak (as to the left) are marked.

$$A_{ij} = \sum a_{i,RT}\ ,\ \ 5.5 + 0.5\,j \le RT < 6 + 0.5\,j\ ,\ \ j = 1 \text{ to } 208 \qquad \text{Eq 3.2.a,}$$

$$A_{ij} = \sum a_{i,RT}\ ,\ \ 5 + j \le RT < 6 + j\ ,\ \ j = 1 \text{ to } 104 \qquad \text{Eq 3.2.b,}$$

$$M_{ij} = \frac{A_{ij}}{A_{i,\max}} \qquad \text{Eq 3.3a,}$$

$$M_{ij} = \frac{A_{ij}}{A_{i,sum}} \qquad \text{Eq 3.3b.}$$

where i and j are indexes for the chromatograms and the time segments, respectively; RT means corrected retention time (section 3.3.1.1); $a_{i,RT}$ is the integrated peak area by retention time RT in the i-th chromatogram; A_{ij} is the sum of peak area values in the j-th time segment of the i-th chromatogram; $A_{i,max}$ is the maximum segment area in the i-th chromatogram; $A_{i,sum}$ is the sum of peak area in the i-th chromatogram; M_{ij} is the normalized value in the i-th row and the j-th column in the measurement matrix.

Conversion results in a matrix containing the samples in rows and the variables in columns, 208 or 104 segment values without additional data extraction.

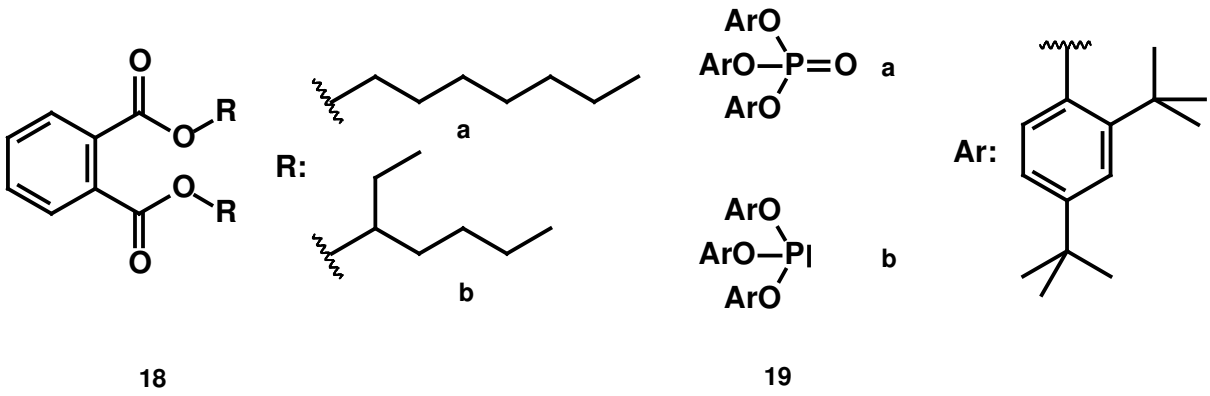
3.3.1.3 Data extraction

It is important to discard samples and variables from the data set that hold no information for clustering. These were empty rows and columns of the measurement matrix and also columns that contained an artifact in almost all samples.

Blank samples, that is derivatizated solvent samples, were measured regularly. These system blanks contained ideally just the solvent peak, thus, together with 'empty' extracts, they were always sorted out automatically during the conversion procedure.

Two strategies were used for the exclusion of redundant variables, called later HCA and PCA/HCA, respectively. As a feature of the program ChromConv, selected columns can be optionally deleted from the measurement matrix during data conversion; then HCA is performed with the remaining data (HCA method). The other process includes feature selection by PCA from the original matrix combined with HCA (PCA/HCA method).

Selection of time segments or variables that do not differentiate the samples is quite challenging. Fermentation blanks, extracts of fermentation mixtures without bacteria, contained many substances in the whole time interval of the GC measurement. Most of the compounds were also present in the bacterial extracts, often in higher concentration than the compounds of interest. Including related time segments in the clustering process results in a clustering according to medium additives instead of substances produced by bacteria.



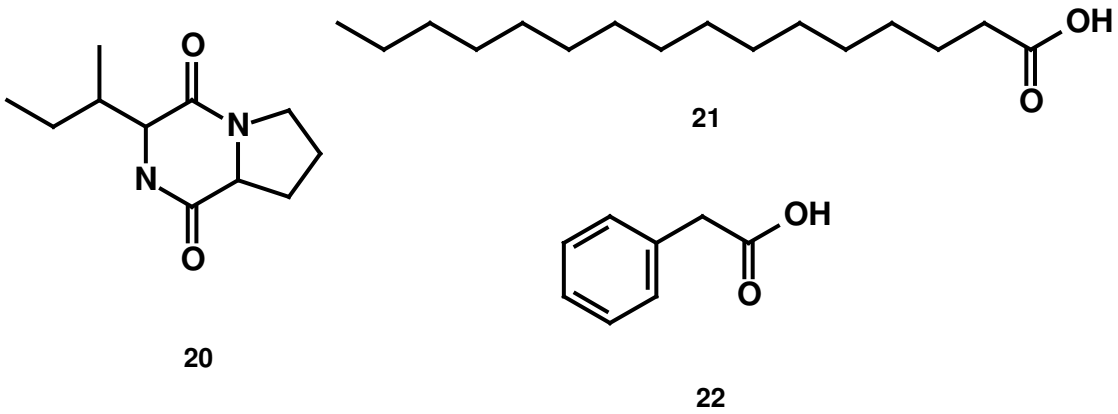
18 19

Scheme 3.3 Artifacts often found in marine extracts: dioctyl phthalate (**18a**), bis-(2-ethylhexyl)-phthalate (**18b**), tris-(2,4-di-tertbutylphenyl)-phosphate (**19a**) and tris-(2,4-di-tertbutylphenyl)-phosphite (**19b**).

The number of substances found in almost every sample grew dramatically, as more and more extracts from control experiments without bacteria were analyzed. Several known artifacts were identified quite early in the experiment. (Scheme 3.3) Peaks of artifacts, such as dioctyl phthalate (**18a**), bis-(2-ethylhexyl)-phthalate (**19b**), tris-(2,4-di-tertbutylphenyl)-phosphate (**19a**) and tris-(2,4-di-tertbutylphenyl)-phosphite (**19b**), can be optionally left out from the clustering process by deleting the related time segment values during data conversion.

Other substances, such as sugars, were already in the fermentation medium as additives. Furthermore, a large group of compounds were often produced during fermentation without the presence of any bacteria. These were diketopiperazines (cyclization products of two amino acids), several fatty acids and aromatic acids, for example. (Scheme 3.4) Since these substances cover altogether a large percentage of the measurement time interval, it was impossible to ignore all the time segments related to them. Nevertheless, segments that involve the peaks of cyc(Ile-Pro) (**20**), cyc(Leu-Pro), and cyc(Phe-Pro) can be optionally discarded during data conversion with ChromConv.

Time segments that contained zero values in more than 95% were deleted from the output file of ChromConv.



Scheme 3.4 Typical components found both in marine extracts and control samples: cyc-(Ile-Pro) (**20**), palmitic acid (**21**), phenylacetic acid (**22**).

3.3.1.4 Overview of the program ChromConv

ChromConv converts GC reports in ASCII format into a two dimensional measurement matrix. Retention time values are corrected and area values are summed up within time segments. Summed area values are organized in a table, where each row represents a sample and each column stands for a time segment of the measurement time interval. The functionality of the program is presented below.

ChromConv

- reads in parameters from a text file and saves them in a hash type of variable;
- modifies certain parameters;
- reads in a text file containing the list of measurements with file names and sample names, and saves this information in a string and a hash, respectively;
- reads the content of the GC directory, opens the GC report files one-by-one, and saves retention time and peak area values in matrixes;
- corrects retention time values and saves them separately in an other matrix;
- sums up peak area values in each time segment for each sample and saves these values in a new matrix;
- changes the segment value to zero for each sample, if the segment contains a peak of an undesired compound (section 3.3.1.3);
- sums up the values in each row of the measurement matrix and saves it in a string type of variable (sum of peak area for each chromatogram);

- finds the maximum segment value for each chromatogram and saves it in a string;
- the relative number of zero values in each segment is saved in a string;
- prints out the final values of the measurement matrix in a text file.

start = 6
end = 110
limit_1 = 315
limit_2 = 375
a_1 = 0.000607
a_2 = 0.000755
b_1 = 0.0982
b_2 = 0.1189
c_1 = 4.4954
c_2 = 5.7489
unwanted_1 = 68
unwanted_2 = 88
unwanted_3 = 91
norm = rel

139|1/ACF,8-3	0	0	0	0	0	0	0	0	3	.
140|1/Bio34	0	0	0	0	0	0	0	50	0	.
141|1/KCy,10	0	1	0	0	59	25	0	0	0	.
142|1/OM13,5	0	0	0	0	0	0	0	0	0	.
144|1/MCy,4	0	0	0	0	0	410	0	3	812	.
145|1//0001	0	2	0	0	0	0	1	77	0	.
146|HKI129-2	0	0	0	0	0	0	0	0	0	.

Fig. 3.6 Part of a text file containing a measurement matrix made by ChromConv. It starts with a list of parameters followed by the matrix itself. The first row of the table, for example, contains the number of a GC file (139), the name of the sample (1/ACF,8-3) and the normalized segment area values.

Output of the program is a table in ASCII format, which contains samples having at least one peak in the chromatogram (beside the solvent peak). Columns that contain more than 95% zero values are not printed out. A part of a typical matrix is presented in Fig. 3.6. Values in the table are normalized within each row according to the related parameter in the parameter file: either to the maximum segment area or to the sum of all areas. Each row starts with a sample identifier, derived from the name of the GC file, followed by the segment values separated by *tabs*. Rows are terminated by *enter*.

3.3.2 Gas chromatographic analysis of marine extracts

3.3.2.1 Derivatization

All the marine extracts were derivatized under standardized conditions before GC analysis. Using derivatives in chromatography increases the volatility of polar compounds and enhances sensitivity due to better peak shape. Compounds produced by bacteria, constituents of cultivation media and artifacts contain polar functional groups (-NH, -OH, -COOH) which otherwise interact with the injector wall and the silica surface of the GC capillary.

The two most widespread micro-scale derivatization methods are trimethylsilylation and methylation. These derivatization reactions are convenient to carry out and the by-products are usually volatile.

Silylation is the method of choice, when derivatization of all protic functional groups is needed in one experiment. It is usually performed in a polar solvent or completely without solvent. Since silylating agents are sensitive to moisture, reactions are carried out in excess of the reagent in sealed vials. Trimethylsilyl (TMS-) derivatives are most widely used and several libraries are available with reference spectra for the fast identification of the compounds. The intensive *m/z* 73, 75 and (M-15) fragments are characteristic for the electron impact ionization mass spectra of the most trimethylsilylated substances (Fig. 3.7).

Marine samples were derivatized with *N*-methyl-*N*-trimethylsilyltrifluoracetamide (MSTFA) for the routine GC analysis. It is a very good silyl donor and its by-products are even more volatile, than that of *N*,*O*-bis(trimethylsilyl)trifluoracetamide (BSTFA).[40] MSTFA was used in excess (section 5.3.1).

Methylation is usually more selective; diazomethane derivatizes only carboxylic -OHs, and trimethylsulfonium hydroxide (TMSH) leaves aliphatic -OHs underivatized[41]. As an important property of TMSH, it is capable of transetherification of glycerine esters.[42] Although the MS fragmentation pattern of methyl esters are much more informative than that of the trimethylsilyl ones, substantial information can be lost via methylation, since marine bacteria produce fatty acid esters themselves.

Although diazomethane and TMSH were used in a later stage for identification purposes, before clustering solely trimethylsilylation was applied.

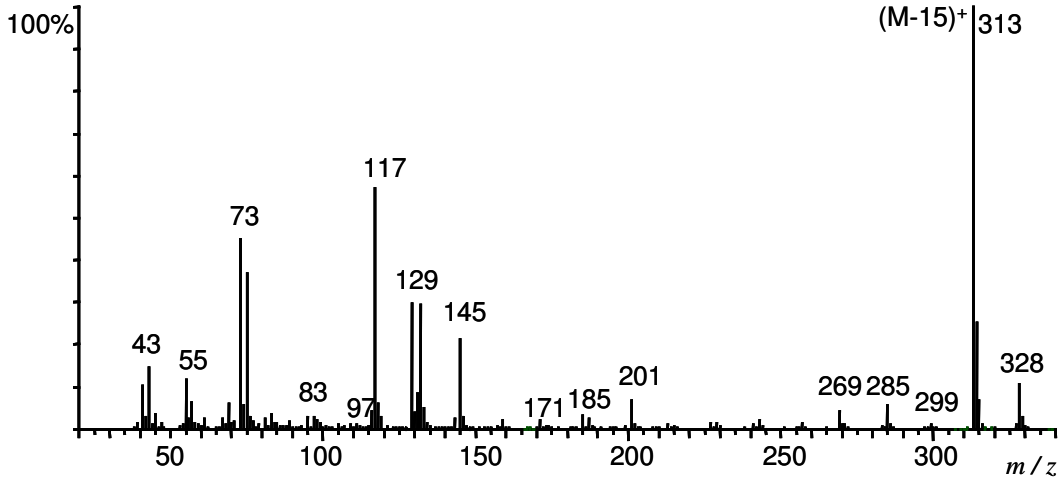


Fig. 3.7 MS spectrum of palmitic acid TMS ester.

3.3.2.2 Gas chromatographic analysis

A separate GC system with a flame ionization detector (FID) was used for the analysis of the marine extracts to avoid impurities of other origin. Moreover, all the measurements were carried out on the same unpolar capillary BPX-5. (section 5.2) Based on the regular analysis of the alkane standard (C11-C44) changes of retention time was followed. A slow GC ramp, 3^{o}C/min was chosen in order to get better separation of the extremely large number of components. Still, overlapping of certain chromatographic peaks could not be avoided owing especially to the tailing peaks of amines and amides, the diketopiperazines, for example. Injections were made in splitless mode to be able to detect trace components.

Although concentration of the samples varied in a relatively large range depending on the scale and yield of fermentation, these differences were successfully omitted in most cases by proper dilution.

Since no internal standard was used during fermentation, it did not make sense to introduce one at this point of the investigation.

3.3.3 Cluster analysis of marine extracts

Two methods were established: hierarchical cluster analysis of GC data after discarding segments of artifacts and diketopiperazines manually (HCA), and data reduction by principal component analysis of the original data set followed by HCA with extracted components as variables (PCA/HCA).

For the PCA/HCA method conversion of GC data was performed with time segments of one minute in order to have less than hundred variables finally. (Eq. 3.2b) Feature selection was made by PCA. The number of components that describe a given percent of the variance in the data set was examined. Using selected factors as new variables HCA was performed with the score values.

The output file of ChromConv is easy to process by SPSS. Once imported, it is saved in a special format. Each row represents a case, a marine extract, and each column stands for a variable, a time segment in the GC chromatogram. That is, samples are clustered according to their segment area values.

The summed area value in a time segment represents a continuous variable, thus the Euclidean and squared Euclidean distances were used to measure similarity between cases. Since concentration or peak area value of the different compounds was in the same scale, no extra scaling of variables was necessary. Normalization to the sum of peak area values within samples allowed to compensate differences in concentration and hence was preferred.

Average linkage was usually applied for merging clusters. Nevertheless, clustering of data sets was also examined with single and complete linkage.

Hierarchical cluster analysis was performed on a specified part of the measurement time interval, when samples containing a given compound had to be found. For this purpose only the related part of the GC data was converted by ChromConv. A particular section of a chromatogram was selected by changing the values ’start’ and ’end’ in the parameter file for ChromConv.

3.4 Results

3.4.1 Clusterification

Since objects are always found to be similar to some extent by clustering, it is advisable to declare a similarity level. Similarity was defined by observation of the GC chromatograms. Three marine extracts were found to have very similar GC chromatograms: 1/IHg,3 and 1/IHg,4 were almost identical, whereas 1/ACF,5 contained the same components but in smaller concentration. The chromatograms are presented in Fig. 3.8. These samples were included in every data set and used as 'similarity standards' due to their unique GC pattern.

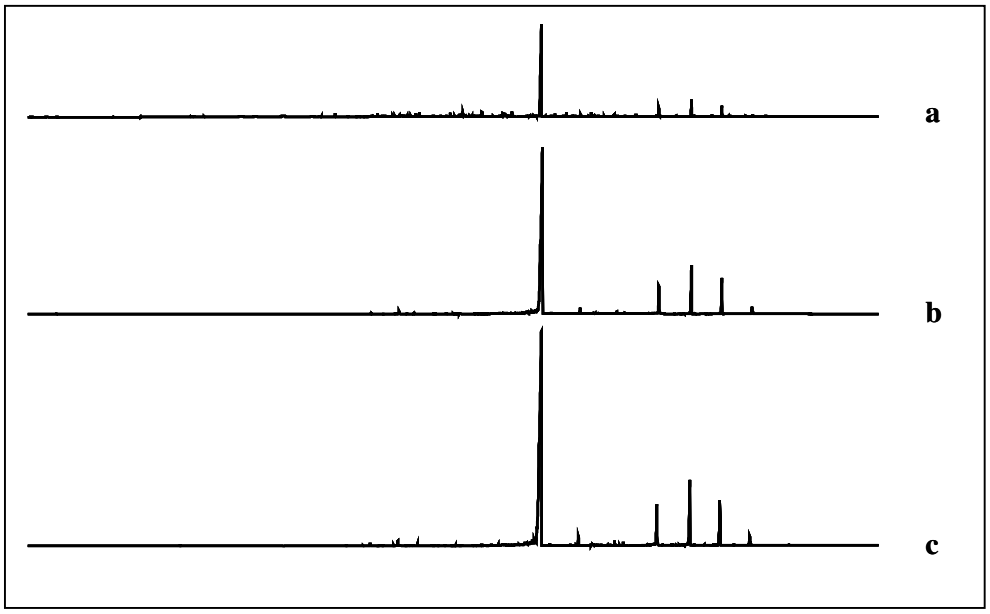


Fig. 3.8 GC chromatogram of the 'similarity standards' **a**) 1/ACF,5, **b**) 1/IHg,4, and **c**) 1/IHg,3.

Samples with the same clustering pattern as seen in the dendogram for 1/IHg,3 and 1/IHg,4 were declared to be very similar. The limit distance for similarity is mostly 4-8 % of the largest distance in the data set.

Several samples were derivatized and analyzed or just analyzed repeatedly to control reproducibility of the method. In case these samples were not together in a cluster, it was more likely due to a changed GC pattern, some impurities in the chromatogram, than some failure in the clustering process.

Squared Euclidean distance measure was preferred over the Euclidean one to calculate distances between objects, since the latter is more sensitive for small differences and our samples were in general very different.
Three linking methods were examined in parallel use of the squared Euclidean distance: single, complete and average linkage. Since the results were independent from the linkage method applied, average linkage between groups was the method of choice.
Principal component analysis was used only for feature selection followed by hierarchical cluster analysis with the score values (PCA/HCA). Clustering of the samples merely by PCA was not successful, the first three principal components represented usually less than 30% of the whole variance.

3.4.1.1 Clustering of lipophilic fractions

The dendogram made by clustering all the extracts was difficult to interpret. Although at the beginning was helpful to find interesting extracts (section 3.4.3), the number of samples reached nearly 500 at the end, which resulted in distorted clusters. Too many samples were found to be similar. Nevertheless, some groups are recognizable. (Dendogram 8.11 in Appendix) It is also interesting to see how different samples from the same bacterial strain were placed in the dendogram; they are mostly 'far' from each other due to different background substances from different fermentation media.
Meaningful clustering was achieved by dividing the whole data set into several subsets, which were analyzed by HCA and PCA/HCA separately. It was based on the experience that conditions have a big influence on the outcome of the fermentation of bacteria. Due to lack of information about composition of fermentation media samples were put into a set by the date of posting from Göttingen, where fermentation was performed. These subsets are called STM2, STM3, STM4, and STM5. A batch of samples called parallel fermentation data set was investigated separately. Data analysis of the different data set is described in the following sections.

Parallel fermentation data set

Several strains, Hel38 (α-Proteobacterium), Hel59 (*Marinomonas* sp.), Hel73 (*Brevibacter linens*), IHg,3 (*Cytophaga* sp.), IHg,4 (*Cytophaga* sp.), Pic006 (*Frigoribacter* sp.) and Pic009 (*Cytophaga* sp.), and control samples without bacteria were fermented in parallel experiments

in four different media: A, C, E, and F. (Table 3.2) Each medium contained half-concentrated artificial sea water.

The fermentation mixture was filtered through celite, extracted with ethyl acetate and concentrated in vacuum. All the extracts were derivatized and analyzed by GC-MS.

Table 3.2 Composition of the fermentation media A, C, E and F in the parallel experiments.

	Concentration (g/l)										
	malt	yeast	flour	fish-meal	glucose	peptone	tryptone	meat extract	$CaCl_2$	NaCl	trace el. sol.*
A	10	4	-	-	4	-	-	-	-	-	-
C	-	1	-	-	10	2	-	1	-	-	-
E	-	5	-	-	5	-	10	-	-	10	-
F	-	10	10	5	21	-	-	-	0.5	-	0.01

* trace element solution

It was interesting to see whether clustering occurred according to strains or fermentation media, and how the control samples clustered. Each time segment of the chromatograms was included in the data analysis procedure to be able to compare control samples with real extracts.

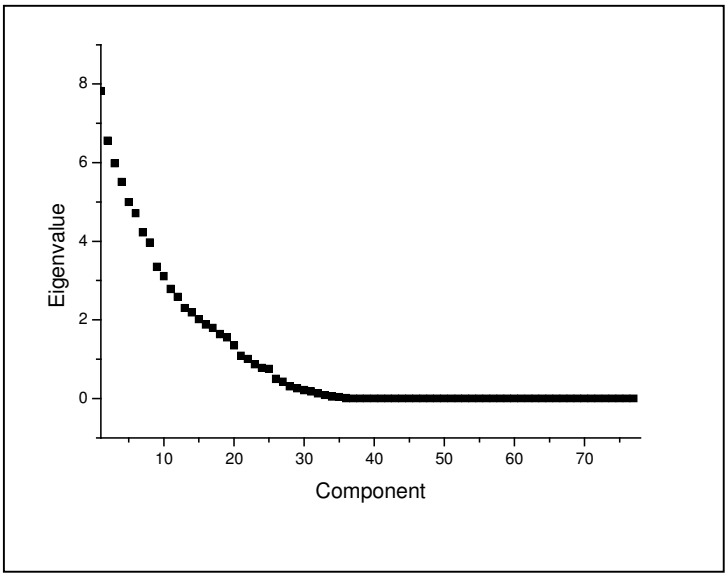


Fig. 3.9 Scree plot resulting from the PCA of the parallel fermentation data set.

Application of PCA for feature selection was examined. By PCA the first 10-20 components covered 65-90 % of the variance. (See scree plot in Fig. 3.9.) Extracting, for example, the first

13 components, 75 % of the variance, and saving the score values in the SPSS table made it possible to perform HCA with these components as new variables. That is, the number of variables was reduced from 77 to 13 in this case.

Results obtained by the methods HCA and PCA/HCA were compared. (Dendogram 8.1 and 8.2 in Appendix.) The IHg,3 and IHg,4 samples in fermentation media A, C and E are clustered together in both cases, just like the samples Hel73 C and Pic006 C. The method PCA/HCA revealed more outliers in the data set, whereas by normal HCA a completely different sample, an alkane standard was also clustered together with a few marine extracts.

These results showed that except the aforementioned examples no clustering of the samples was meaningful. Marine extracts in this data set were not significantly different from control samples. Moreover, the difference in the composition of the fermentation media (A, C, E, F) had a negligible influence on the GC pattern of the samples.

Main components found in the bacterial extracts were fatty acids and sugars. Nevertheless, these compounds were background substances, also present in control samples. Table 3.3 represents typical components identified by GC-MS in control samples.

Table 3.3 Substances found in control samples from parallel fermentation. -: not detected, +++: 50-100%, ++: 10-50 %, +: 0-10%, (+): below limit of calculation (% of the maximum peak area).

	Medium A	Medium C	Medium E	Medium F
2-phenylethyl alcohol	+	+	(+)	-
2-(2-hydroxyphenyl)-ethylalcohol	++	++	-	++
benzoic acid	-	-	+	+
phenylacetic acid	+	-	+	-
2-hydroxyphenylacetic acid	-	-	+	-
3-phenylpropionic acid	-	-	-	++
2-(3-indole)ethanol	+	+	-	+
3-indoleacetic acid	-	-	+	-
bis-(2-ethylhexyl) phthalate	+	+	+	+
3ß-hydroxy-5ß-cholestane	-	-	+	-
3ß-hydroxy-5-cholestene	-	-	++	++
3ß-hydroxy-24-ethylcholest-5-ene	-	-	-	+
cyc(Ile-Pro) (2 isomers)	-	+	+	+
cyc(Leu-Pro) (2 isomers)	-	+	+	+
cyc(Phe-Pro) (2 isomers)	-	+	+	(+)
cyc(Pro-Tyr) (2 isomers)	+	+	-	-
glycerin	+	+	+	+
monomyristin	-	-	-	+
glycerin monopentadecanoate	-	-	-	+
monopalmitin	+	-	+	-
glycerin monolinoleate	-	-	+	+
glycerin monooleate	-	-	+	-
monostearin	-	-	+	+
2-keto-D-gluconic acid	+	++	+	++
sorbose/fructose	-	+	+	+
ß-D-galactofuranose	-	++	+	+
α-D-glucopyranose	++	+++	+++	+++
mannose	+	+++	+++	+++

Table 3.3 Continued.

	Medium A	Medium C	Medium E	Medium F
decenoic acid	+	+	-	-
capric acid	+	+	-	-
lauric acid	+	+	-	-
12-methyltridecanoic acid	+	-		
11-methyltridecanoic acid	+	-	+	++
13-methylmyristic acid	+	-	-	+
12-methylmyristic acid	+	-	+	(+)
pentadecanoic acid	+	-	-	+
hexadecenoic acid	+	-	+	+
14-methylpentadecanoic acid	+	-	+	+
13-methylpentadecanoic acid	-	-	++	+++
palmitic acid	++	+	-	-
heptadecenoic acid	-	-	-	+
15-methylpalmitic acid	+	-	-	-
14-methylpalmitic acid	+	-	+	-
heptadecanoic acid	-	-	+	+
Z,Z-9,12-octadecenoic acid	+++	-	++	+++
Z-11-octedecenoic acid	+	-	++	+++
stearic acid	+	-	+	+
arachidic acid	-	-	-	+
hexadecyl alcohol	-	+	-	-
octadecyl alcohol	-	+	-	-

Subset STM2

This was the first batch of samples obtained in 2000 containing 98 extracts. Composition of fermentation medium (or media) was unknown. HCA was performed after deleting undesired segments by ChromConv (Dendogram 8.3 in Appendix). By the other method 25 components were extracted from 75 by PCA, which represented 79 % of the variance, and HCA was made with the score values. (Dendogram 8.4 in Appendix.) The two dendograms were compared. Several groups of similar samples were found with both methods.

A few samples were further analyzed by GC-MS. Some extracts in one of the clusters contained two isomers of oleic amide as main components. The cluster of 1/IHg,3, 1/IHg,4, and 1/ACF,5 (probably *Cytophaga* sp.) (Fig. 3.10) was further investigated by HR-MS and other methods, described in section 3.4.3.1. In fact, by the combination of PCA and HCA

another sample, 1/OM13,5 (*Halomonas* sp.) was found to contain the same series of compounds.

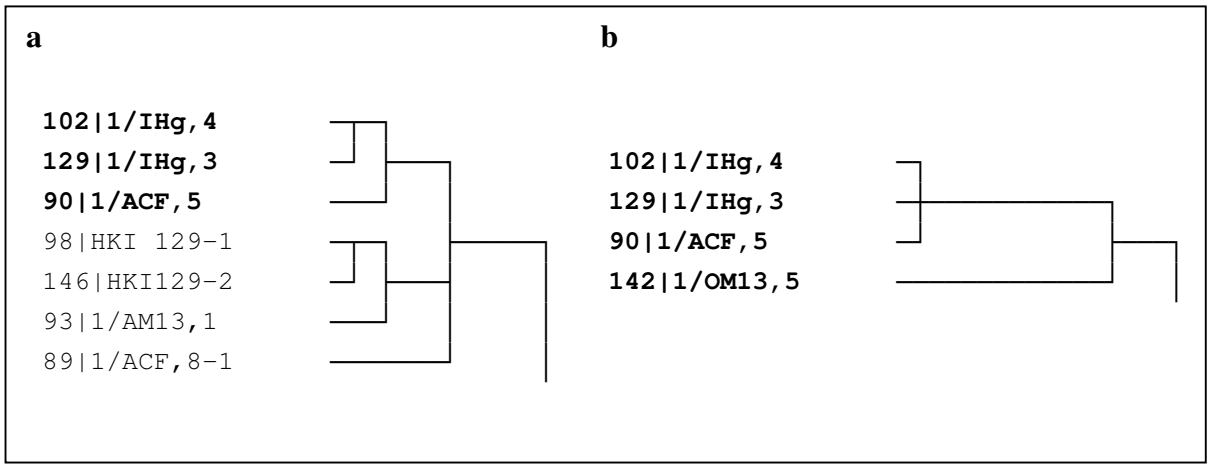


Fig. 3.10 Part of the dendogram of sub data set STM2 made **a**) by HCA **b**) PCA/HCA. A forth extract, 1/OM13,5 was found by method **b** to contain the same substances as the other marked samples. These samples have a very typical GC pattern.

Novel terpenes **26** and **27** and isomers of a lactone (discussed in section 3.4.3.4) were found in the extract of the strain 1/AMP,3 (*Cytophaga* sp.). Isolation and structure elucidation of the terpenes is presented in section 3.4.2.

The sample 1/Hel26 (*Roseobacter* sp.) contained amides of palmitic and oleic acids, and the *iso*-propyl ester of palmitic acid.

Subset STM3

This batch, obtained in 2000, contained 91 samples. Composition of fermentation medium (or media) was unknown. Dendograms were made again by both methods: HCA (Dendogram 8.5 in Appendix) and PCA/HCA . By the second method 20 components were extracted from the original 70 variables by PCA (74 % of variance) and HCA was performed with the score values. (Dendogram 8.6 in Appendix.) The scree plot from PCA did not have a sharp elbow (see Fig. 3.11), which means the variables were already uncorrelated, or there was a high noise in the data set.

Selected samples from subset STM3 were analyzed by GC-MS. Methylene-1,1'-bis-piperidine was found in a large number of extracts. In sample ANT253(RE) a series of *N*-acetyl amides was identified by GC-MS. (see Scheme 3.5) Two extracts, ANT323 and ANT46 were found to have almost the same GC pattern as the 'similarity standards'

(introduction to section 3.4.1), but the substances identified by GC-MS turned out to be esters of long-chain fatty acids and long-chain alcohols.

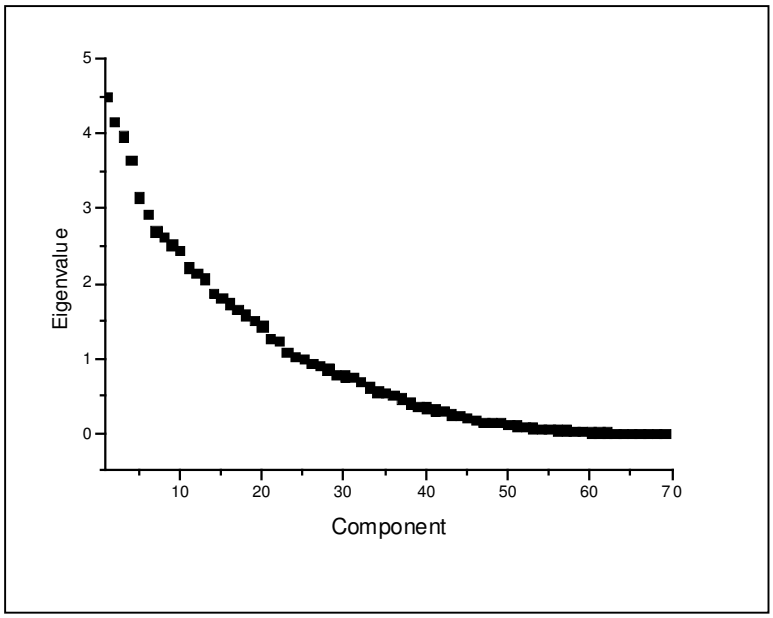
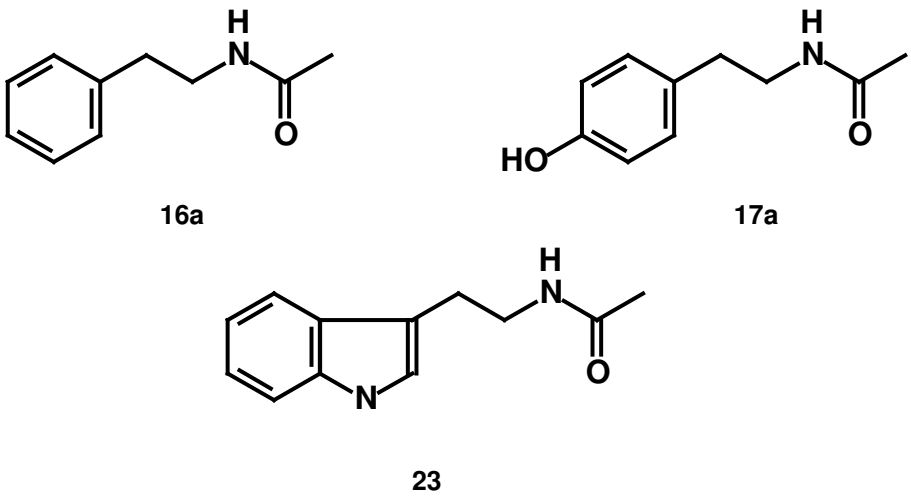


Fig. 3.11 Scree plot made by PCA of the samples in subset STM3; no sharp elbow can be observed.



16a **17a**

23

Scheme 3.5 *N*-acetyl amides from ANT253(RE). The amides **16a** and **17a** were also identified in marine *Streptomyces* strains in the first period of the project (section 3.1).

The sample VH Hel10 (*Roseobacter* sp.) contained alkanes, long-chain alcohols, fatty acids, and 2-(indo-3-yl)-ethanol.

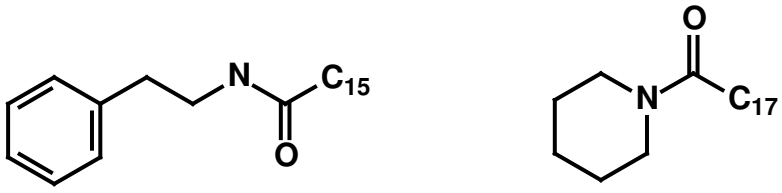
Subset STM4

This batch of 76 samples was obtained in April 2001. The composition of the fermentation medium (or media) was unknown. HCA was made after deleting undesired segments with ChromConv. (See Dendogram 8.7 in Appendix.) By the subsequent use of PCA and HCA all segments were included and 20 components (from 88) were extracted, which represented 74% of the variance. HCA was made with the score values, as usual. (See Dendogram 8.8 in Appendix.) Samples from control derivatization and injection were clustered mostly together.

Certain extracts from the data set were selected to be analyzed by GC-MS. A group of strains were found to be producers of *N*-(phenylethyl)-acetamide (**16a**) in larger amounts. A long-chain derivative **24** was identified by GC-MS in Pic006 (Scheme 3.6). In fact, *N*-acetyl derivatives of other amines were also common in the data set.

The strains Pic006 (*Frigoribacter* sp.), Hel38 (α-proteobacterium), Bio137 (*Cytophaga* sp.) and Bio138 (probably *Cytophaga* sp.) inhibit the growth of specific tumor cells, such as gastric, liver, and breast cancer. Long-chain ketones with (n-1)/(n-2) methyl branching were identified by GC-MS and other techniques in Pic009 and Pic006, presented in section 3.4.3.2.

In the extracts Bio137 and Bio138 sulphur heterocycles were found, summarized in section 3.4.3.3. Hel38 contained squalene and the C35 terpenes **26** and **27** (section 3.4.2) beside aromatic acides, amines, and long-chain amides.



24 **25**

Scheme 3.6 Long-chain *N*-acyl derivative **24** from Pic006 (subset STM4) and a long-chain *N*-acyl derivative of piperidine, **25,** from Hel45 and Hel73 (subset STM5).

The cyclohexane extract of Hel45 (*Roseobacter* sp.) contained antranilic acid, antranylamide, indole and its derivatives, fatty acid methyl esters, and ß-hydroxy-capric acid.

Subset STM5

This batch of 67 samples was obtained in July 2001. The composition of the fermentation medium (or media) was unknown. HCA was performed after deleting undesired segments

with ChromConv. (See Dendogram 8.9 in Appendix.) By PCA 20 components were extracted from 80 variables, which was 76 % of the variance, then HCA was made with the score values. (See Dendogram 8.10 in Appendix.) Repeatedly injected samples were clustered together more or less by both methods.

A cluster of extracts was revealed containing the C35 terpenes **26** and **27**, described in section 3.4.2, and it was more clearly separated by PCA/HCA. (Fig. 3.12) Closer examination of the results of the GC and GC-MS analysis led to the discovery that all these samples contained diketopiperazines in a high concentration, which was the reason for the clear clustering by PCA/HCA.

The strains Hel10, Hel26 and Hel45 (all *Roseobacter* sp.) were fermented for the second time (see samples VH Hel10, 1/Hel26 and Hel45 above). New substances were the C35 terpenes **26** and **27** identified in Hel10 and Hel45 and the amide **25** from Hel45 and Hel73 (*Brevibacter linens*). (Scheme 3.6)

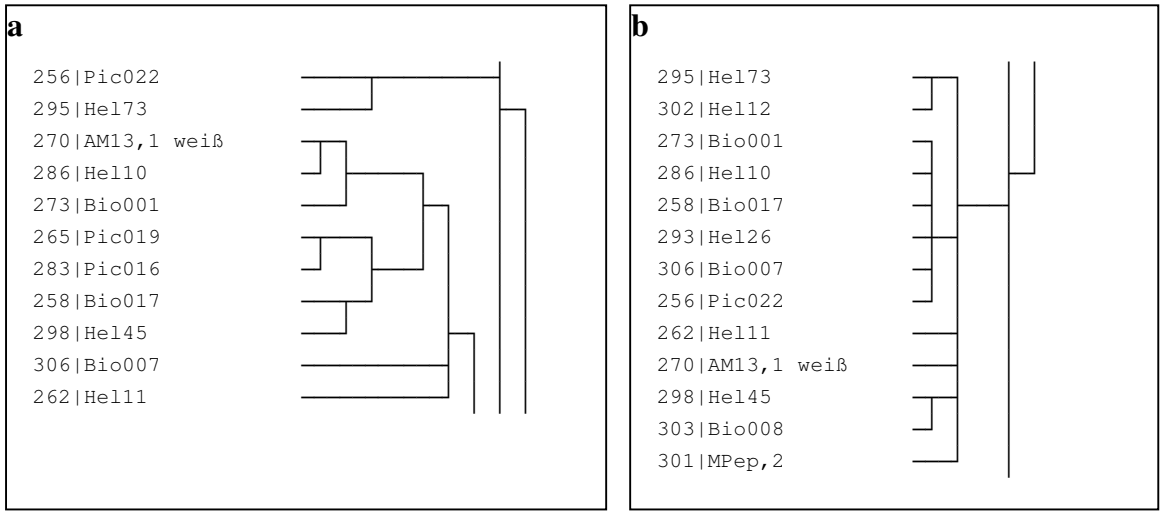


Fig. 3.12 Cluster of the C35 terpenes in the sub data set STM5 made by HCA (**a**), and PCA/HCA **(b**).

Hel73 turned out to be another producer of the long-chain ketones with (n-1)/(n-2) methyl branching (see section 3.4.3.2).

3.4.2 Isolation and structure elucidation of two novel C35 terpenes

A pair of late eluting peaks was observed in the chromatogram of AMP,3 (*Cytophaga* sp.) with a molar mass of 474 and 476, and a retention index *RI* = 3363 and 3368, respectively.

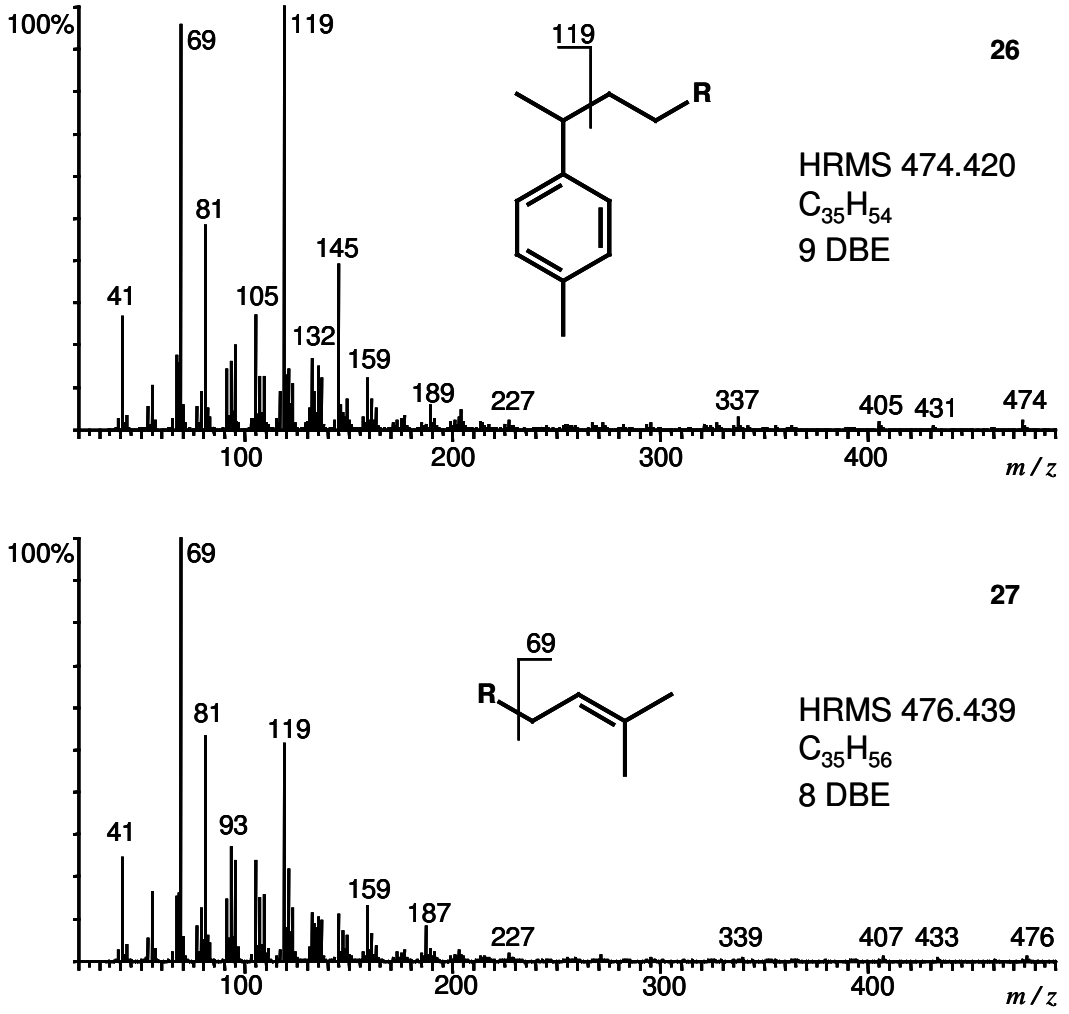


Fig. 3.13 MS spectra of the two terpenes **26** and **27** with HRMS data.

The MS spectra of the substances were terpene-like and similar to that of *ar*- and β-curcumene.[43] The *m/z* 69 fragment is typical for oligoprenyl chains. (Fig.3.13) The main fragment of the terpene **26** (M474), the *m/z* 119 derives from the cleavage of the single bond at the methyl branching nearest to the aromatic ring. (Fig.3.13) GC-HRMS analysis revealed the molecular formula of the substances **26** and **27**: $C_{35}H_{54}$ and $C_{35}H_{56}$ with 9 and 8 double bond equivalents, respectively.

Cluster analysis with segments from the particular part of the GC chromatograms, where the peaks eluted, led to the discovery of other samples that contained the two terpenes.

(Dendogram 8.12 in Appendix) In order to obtain more structural information, the substances were isolated.

3.4.2.1 Liquid chromatographic isolation

Samples that contained the terpenes in a higher concentration were chosen for the isolation based on the results of the cluster analysis. More material was obtained by additional fermentation of AMP,3 and Hel59 in a larger scale (5 L). The lipophilic fraction of the samples was dissolved in n-pentane, loaded onto a silica column and eluted with n-pentane. The terpenes were separated from more polar compounds, but not from squalene. Several fractions contained both terpenes.

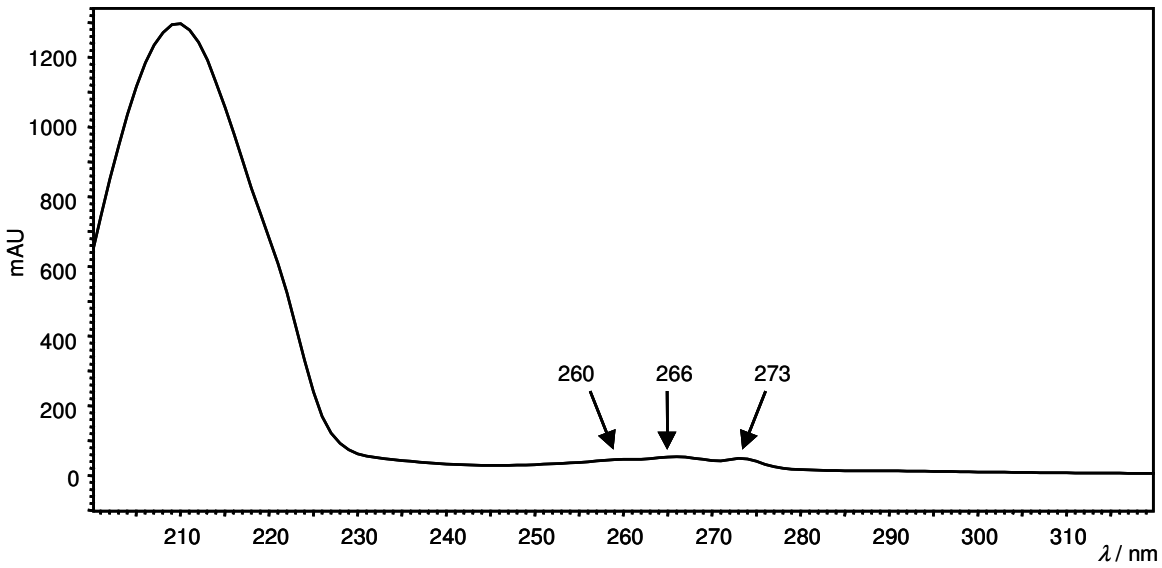


Fig. 3.14 UV scan from the HPLC peak of **26** at 15.78 min

The fractions of interest were united and further purified by NP-HPLC. The best separation of the three compounds was obtained on a Supersphere Si60 column with n-hexane as a mobile phase (see section 5.2 for details). The UV detection was at 215 nm in the absorption range of polyenes. Small maxima at 266, 260 and 273 nm could be observed in the UV spectrum of **26**, which was a hint for the aromatic character of the substance. This way 1-1.5 mg of **26** and approx. 0.5 mg of **27** were isolated.

3.4.2.2 Determination of the number of double bonds by catalytic hydrogenation

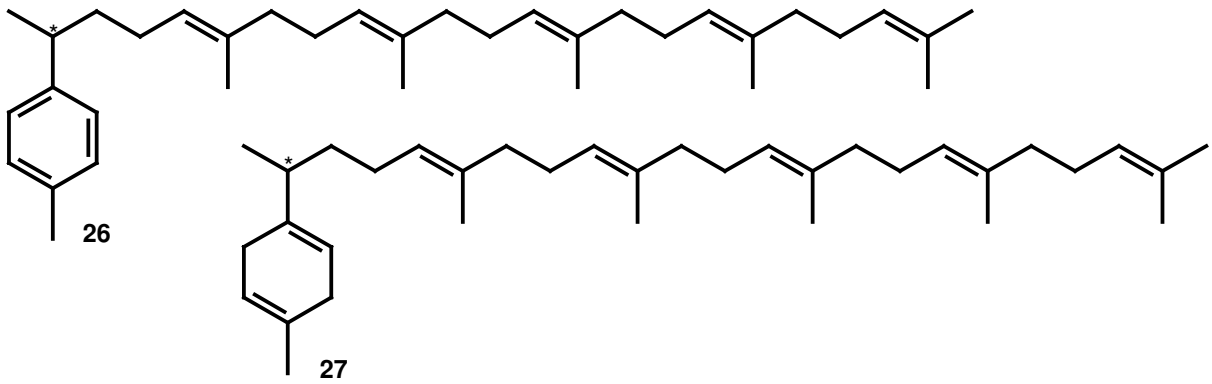
The fraction containing the terpene **27** was hydrogenated with Pd/C to determine the number of aliphatic double bonds. The terpene M476 was already partly isomerized to M474, which could be observed from the GC-MS analysis made directly before hydrogenation. As a result, a mixture of partially saturated compounds was obtained with a molar mass of 480, 482, 484 and 486. All these products contained the aromatic ring except for the M486. Consequently, both terpenes possess five aliphatic double bonds and a ring, which is aromatic in **26**. That correlated again with the structure of the curcumenes.

Substance **27** was analyzed by nuclear magnetic resonance spectroscopy (NMR) before hydrogenation and partly aromatized during these experiments.

3.4.2.3 Structure elucidation by NMR spectroscopic experiments

The isolated terpenes **26** and **27** were analyzed by NMR (400 MHz) in different solvents. Due to the small amount of the samples, it was not possible to carry out ^{13}C-NMR experiments. However, the chemical shift values of the carbon atoms of compound **26** could be observed by H,C-correlation measurements (HSQC, HMBC).

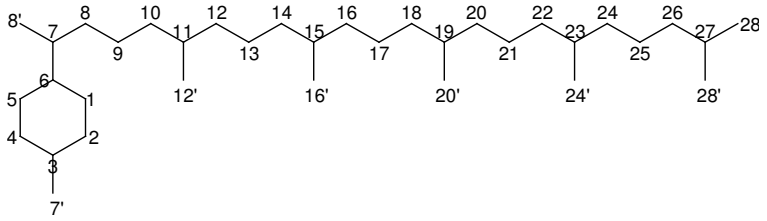
^{1}H-NMR and HSQC spectra from **26** and ^{1}H-NMR spectrum from **27** were recorded first in deutero-dichloromethane. The large peak at 1.52 ppm in both of the ^{1}H-NMR spectra revealed the presence of a water peak, which was overlapping a signal of methyl protons (16.1 ppm).



Scheme 3.8 Structure of the C35 terpenes **26** and **27**.

The attempt to remove water from the sample containing **27** in vacuum was only partly successful; although the water peak became smaller, other impurities (ether) were detected in the sample.

The proton signal of water in deutero-benzene does not disturb the methyl signals, since it lies at 0.40 ppm. Therefore, deutero-benzene was chosen as a solvent for further experiments: ^{1}H-NMR of terpene **27**, and HSQC and HMBC of terpene **26**.



Scheme 3.7 Numbering of the atoms corresponding to NMR data in Table 3.4, 3.5, and 3.6.

Table 3.4 Comparison of the ^{1}H-NMR data of the terpene **26** and that of *ar*-curcumene.

Position of the protones		^{1}H-NMR chemical shifts in ppm	
ar-curcumene	**26**	*ar*-curcumene in $CDCl_3$[44]	**26** in CD_2Cl_2
8’	8’	1.19 (d, 3H)	1.20 (d, 3H)
12’	12’ 16’ 20’ 24’ 28’	1.50 (s, 3H)	1.52 (s, 3H), 1.59 (s, 12H)
8	8 12 16 20 24	1.51-1.65 (m, 2H)	1.82-2.00 (m, 10H)
12	28	1.65 (s, 3H)	1.67 (s, 3H)
9	9 13 17 21 25	1.81-1.89 (m, 2H)	2.02-2.10 (m, 10H)
7’	7’	2.30 (s, 3H)	2.29 (s, 3H)
7	7	2.60-2.67 (m, 1H)	2.60-2.70 (m, 1H)
10	10 14 18 22 26	5.04-5.10 (m, 1H)	5.07-5.14 (m, 5H)
1 2 4 5	1 2 4 5	7.04 (s, 4H)	7.07 (dd, 4H)

The ^{1}H-NMR spectra and the ^{13}C signals obtained by HSQC and HMBC for **26** match the corresponding signals of *ar*-curcumene[43] (presented in Table 3.4 and 3.6). H,C-correlations are presented in Scheme 3.9 and 3.10.

Table 3.5 Comparison of the ^{1}H-NMR data of the terpene **27** and that of β-curcumene.

Position of the protones		^{1}H-NMR chemical shifts in ppm	
β-curcumene	**27**	β-curcumene in $CDCl_3$[43]	**27** in CD_2Cl_2
8’	8’	0.99 (d, 3H)	0.98 (d, 3H)
12’	12’ 16’ 20’ 24’ 28’	1.20-1.60 (m, 5H)	1.58 (s, 3H), 1.60 (s, 12H)
8	8 12 16 20 24		1.87-2.02 (m, 10H)
7’ 12	7’ 28	1.67 (br.s, 6H)	1.65 (s. 3H), 1.67 (s, 3H)
9 7	9 13 17 21 25 7	1.90 (m, 2H), 2.09 (m, 1H)	2.02-2.13 (m, 11H)
2 5	2 5	2.58 (m, 4H)	2.56 (br.s, 4H)
10	10 14 18 22 26	5.09 (m, 1H)	5.06-5.15 (m, 5H)
1 4	1 4	5.42 (s, 2H)	5.42 (m, 2H)

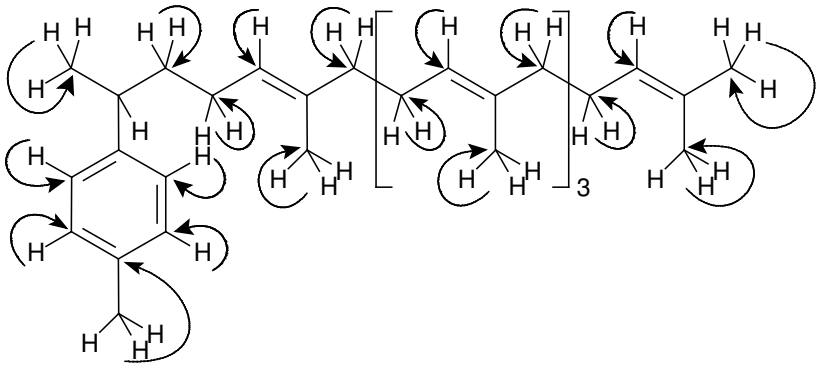
Table 3.6 Comparison of the ^{13}C-NMR data of the terpene **26** and that of *ar*-curcumene. ([a]from HMBC).

Position of the carbon atoms		^{13}C-NMR chemical shifts in C6D6 in ppm	
ar-curcumene	M474	*ar*-curcumene[45]	**26**[a]
12’	12’ 16’ 20’ 24’ 28’	17.7 (q)	16.1, 16.2 (3C), 17.8
7’	7’	21.0 (q)	21.1
8’	8’	22.9 (q)	23.1
12	28	25.8 (q)	25.8
9	9 13 17 21 25	26.6 (t)	27.2 (5C)
8	8 16 20 24 12	38.8 (t)	38.9, 40.2 (3C), 40.3
7	7	39.4 (d)	39.5
10	26 14 18 22 10	125.1 (d)	125.0, 124.7-124.9 (3C), 125.2
1 5	1 5	127.3 (d, 2C)	no (H,C) corr. observed
2 4	2 4	129.3 (d, 2C)	129.4 (2C)
11	27 11 15 19 23	131.1 (s)	131.2, 135.0-135.1 (4C)
3	3	135.3 (s)	135.3
6	6	144.7 (s)	144.9

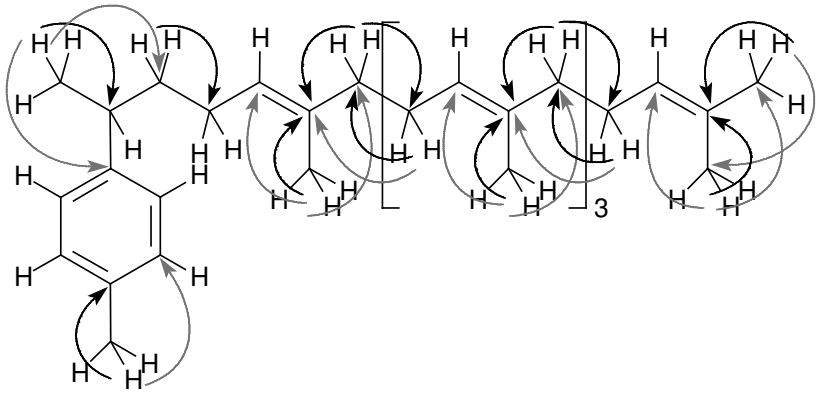
The difference in the number of methyl groups between 1.5-1.6 ppm (5 instead of 1), the number of methylene groups (10 instead of 2), and the number of tertiary protons between 5.0-5.2 ppm (5 instead of 1) demonstrates that **26** contains four more prenyl units than *ar*-curcumene.

Likewise, the similarities (and differences) of the ^{1}H-NMR spectrum of **27** and that of β-curcumene (Table 3.5) showed that this terpene had the same ring structure as β-curcumene,

but again, its chain is four prenyl unit longer. With only eight double bond equivalents, it contains a hexadiene ring.



Scheme 3.9 H,C-correlation in the molecule of **26** from the HSQC experiment.



Scheme 3.10 H,C-correlation in the molecule of **26** from the HMBC experiment, black are the ^{2}J(C,H) and gray are the ^{3}J(C,H) couplings.

Based on the results of the NMR experiments and spectral analogy to the curcumenes the structure of the two terpenes was elucidated (Scheme 3.8). We propose for the new terpenes **26** and **27** the names tetraprenyl-*ar*-curcumene and tetraprenyl-β-curcumene, respectively.

3.4.2.4 Determination of the absolute configuration of 26

Both terpenes contain a chirality center at the position seven. (scheme 3.8) Regarding the extreme small amount of substance available, only chiral gas chromatographic methods were suitable for the determination of the absolute configuration.

Since the low volatility of the terpenes and the long chain makes a direct chiral GC analysis impossible center, the molecule had to be chemically degraded. The aromatic ring was cleaved to obtain the smallest degradation product that still contains the asymmetric carbon.

Several oxidative degradation methods that cleave a double bond are suitable for opening an aromatic ring. As final stage by these methods a carboxylic acid is generated at the position of the cleavage. This degradation offers some advantages; the product is stable and easy to derivatize to the more volatile methyl ester for the chiral GC analysis.

Oxidative degradation of the aromatic ring with ruthenium tetroxide

Cleavage of a resonance stabilized aromatic ring is difficult. The bacterial enzymatic cleavage of an aromatic ring can be oxidative under aerobe conditions or reductive without the presence of oxygen.[46] The chemical methods developed in the lab are oxidative: degradation with trifluoroperacetic acid[47], ozone[48,49] or ruthenium tetroxide[50]. The first mentioned method did not become wide-spread because of the difficulty with handling excess amounts of the reagent.[51,47] Although ozonolysis could be successful with ozone concentrations more than 15%,[48] the reaction itself and the work-up methods are not really suitable for micromolar levels.

Alternatively, an oxidation method using ruthenium tetroxide, generated continuously in the reaction mixture with a stoichiometric oxidant (mostly sodium periodate) in a biphasic system, proved to be extreme useful in the natural product research.[52,53] Alkyl-substituated benzenes are first oxidized to an α-keto-aldehyde and four molecules of formic acid, which is further oxidized to carbon dioxide.[54] The reaction of the keto-aldehyde with ruthenic(VIII) acid (or the aldehyde-hydrate with ruthenium(VIII)-tetroxide) results in the formation of a diester of ruthenic(VIII) acid. Finally, the shift of three valence electron pairs leads to the generation of an alkyl-substituated acid and carbon dioxide.

The mechanism of the attack of ruthenium tetroxide on the olefinic double bond is discussed in detail elsewhere.[55]

The first attempts date back in the early 1950s, when ruthenium tetroxide was used as a substitute oxidizing agent for the strongly toxic osmium tetroxide.[56] Since it is good soluble in carbon tetrachloride, the reaction had been carried out in a two-phase mixture with water.[57] The most cited publication in this topic is from 1981, from Sharpless et al.[50]. They managed to optimize the reaction conditions and found out that adding acetonitril to the mixture helps to maintain the activity of the Ru(VIII) catalyst. A systematic study appeared in 1990[58] revealed the usefulness of periodic acid instead of periodate, especially with small amounts of starting material. The authors also reported about the temperature dependence of the reaction. The volatility of ruthenium tetroxide and the danger of over-heating set an upper limit of approx. 40°C, whereas under 20°C the reaction would be too slow. The catalyst can be deactivated with an excess of dichloromethane[50] or diethyl ether[57,58], which are also suitable solvents for work-up. Electron-donating substituents on the aromatic ring have an activating effect.[57]

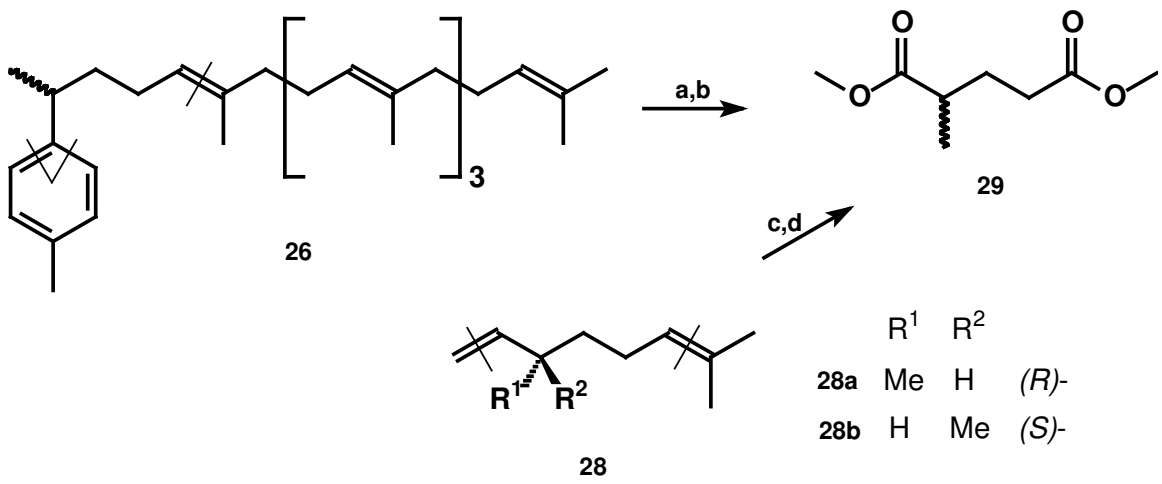


Fig. 3.11 Degradation of citronellene (**28**) and the natural product (**26**) to 2-methylglutaric acid (**29**); **a**) 35 equiv. H_5IO_6 and 2 mol% $RuCl_3.(H_2O)_n$ in $CH_3CN/CCl_4/H_2O$ 2/2/3, **b**) diazomethane; **c**) 8.1 equiv. $NaIO_4$ and 2 mol% $RuCl_3.(H_2O)_n$ in $CH_3CN/CCl_4/H_2O$ 2/2/3, **d**) Amberlyst 15 in methanol.[59]

The natural product **26** was degraded with periodic acid as stoichiometric oxidant and ruthenium tetroxide as catalyst. The micro-scale work-up with dichloromethane was followed by derivatization with diazomethane (see section 5.3 for details). (Scheme 3.11.) As synthetic references, both enantiomers of citronellene were oxidized to 2-methylglutaric acid.

Chiral separation

The *R* and *S* enantiomers of dimethyl 2-methylglutarate (**29a** and **29b**) could be separated on a Hydrodex-6-TBDMS (15m, 0.25mm) column by 60°C isotherm and a linear velocity of 67 cm/s (hydrogen as carrier). The peak resolution was 0.7.

The absolute configuration of the degradation product, dimethyl 2-methylglutarate (**29**), was determined by analyzing both synthetic enantiomers and the 1:1 mixture of them by GC-MS on the capillary used for the GC separation. The results were compared to that of the natural product. (See Fig. 3.15.) Identification by MS compensated for the unsatisfactory separation on the same column installed in the GC-MS instrument. Thus the absolute configuration of the original natural **26** product was proved to be *(R)*.

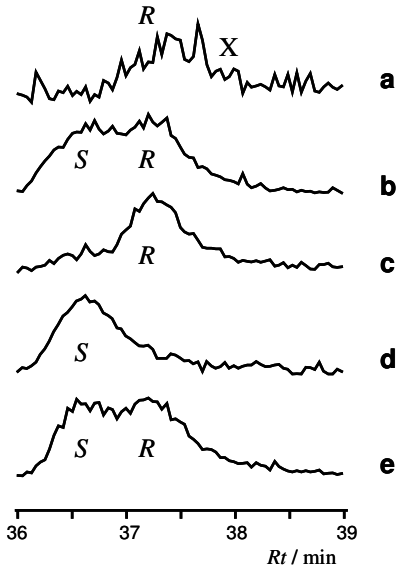


Fig. 3.15 GS-MS separation of the enantiomers of dimethyl 2-methylglutarate. **a**) ion trace *m/z* 114 of the degradation product of the isolated natural product, **b**) ion trace *m/z* 114 of the coinjection with 1:1 mixture of the synthetic enantiomers, **c**) synthetic enantiomer (*R*), **d**) synthetic enantiomer (*S*), **e**) 1:1 mixture of the synthetic enantiomers.

3.4.3 Selected substances from marine extracts

3.4.3.1 Hexanetriol esters from IHg,3, IHg,4, and ACF,5

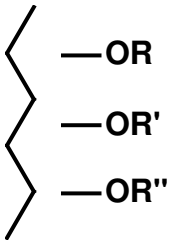
A significant cluster of three samples (Fig. 3.10), IHg,3, IHg,4, and ACF,5 (all *Cytophaga),* were examined. The strains were isolated from cultures enriched from filtered water sample (5 μm pore size) taken from the North Sea.

The GC-MS analysis of the MSTFA derivatized samples revealed a series of structurally related esters ,**30**, **31**, **32**, **33**, **34**, and **35**, containing fragment ions *m/z* 127 and *m/z* 155. The MS spectra of the earlier eluting two compounds contained the fragment *m/z* 73 but not *m/z* 147, which meant that only one functional group was trimethylsilylated. Related $(M\text{-}15)^+$ fragments were *m/z* 471 and *m/z* 499 for the silylated **30** and **31**. The molecular ion of the underivatized compounds was not observable.

Table 3.7 Results of the GC-HRMS analysis.

	Exact mass/ amu	Fragment	Formula	Post. Mol. Form.
m/z 127	127.113	α (acyl)	$C_8H_{15}O^+$	-
m/z 155	155.144	α (acyl)	$C_{10}H_{19}O^+$	-
30-TMS	471.350	$(M\text{-}CH_3)^+$	$C_{26}H_{51}O_5Si^+$	$C_{27}H_{54}O_5Si$
31-TMS	499.382	$(M\text{-}CH_3)^+$	$C_{28}H_{55}O_5Si^+$	$C_{29}H_{58}O_5Si$
32	369.301	$(M\text{-}OC(O)C_7H_{15})^+$	$C_{22}H_{41}O_4^+$	$C_{30}H_{56}O_6$
33	397.329	$(M\text{-}OC(O)C_7H_{15})^+$	$C_{24}H_{45}O_4^+$	$C_{32}H_{60}O_6$
34	397.329	$(M\text{-}OC(O)C_9H_{19})^+$	$C_{24}H_{45}O_4^+$	$C_{34}H_{64}O_6$
35	425.364	$(M\text{-}OC(O)C_9H_{19})^+$	$C_{26}H_{49}O_4^+$	$C_{36}H_{68}O_6$

IHg,3 was further analyzed by a gas chromatograph-high resolution mass spectrometer (GC-HRMS) system. (Table 3.7.) Knowing the exact mass of the fragments of the derivatized and underivatized compounds a general structure was postulated (see Scheme 3.12). These compounds are probably esters of a hexanetriol, acylated with octanoic and/or decanoic acid. The intensity ratio of the acyl fragments, *m/z* 127 and *m/z* 155 in the MS spectrum revealed the acylation pattern (see for example for **33** and **34** in Fig. 3.16).



	R	R'	R''
30	—C(O)C_7H_{15}	—C(O)C_9H_{19}	—H
31	—C(O)C_9H_{19}	—C(O)C_9H_{19}	—H
32	—C(O)C_7H_{15}	—C(O)C_7H_{15}	—C(O)C_7H_{15}
33	—C(O)C_7H_{15}	—C(O)C_7H_{15}	—C(O)C_9H_{19}
34	—C(O)C_7H_{15}	—C(O)C_9H_{19}	—C(O)C_9H_{19}
35	—C(O)C_9H_{19}	—C(O)C_9H_{19}	—C(O)C_9H_{19}

Scheme 3.12 The general structural formula postulated for the hexantriols and the R groups of the particular substances.

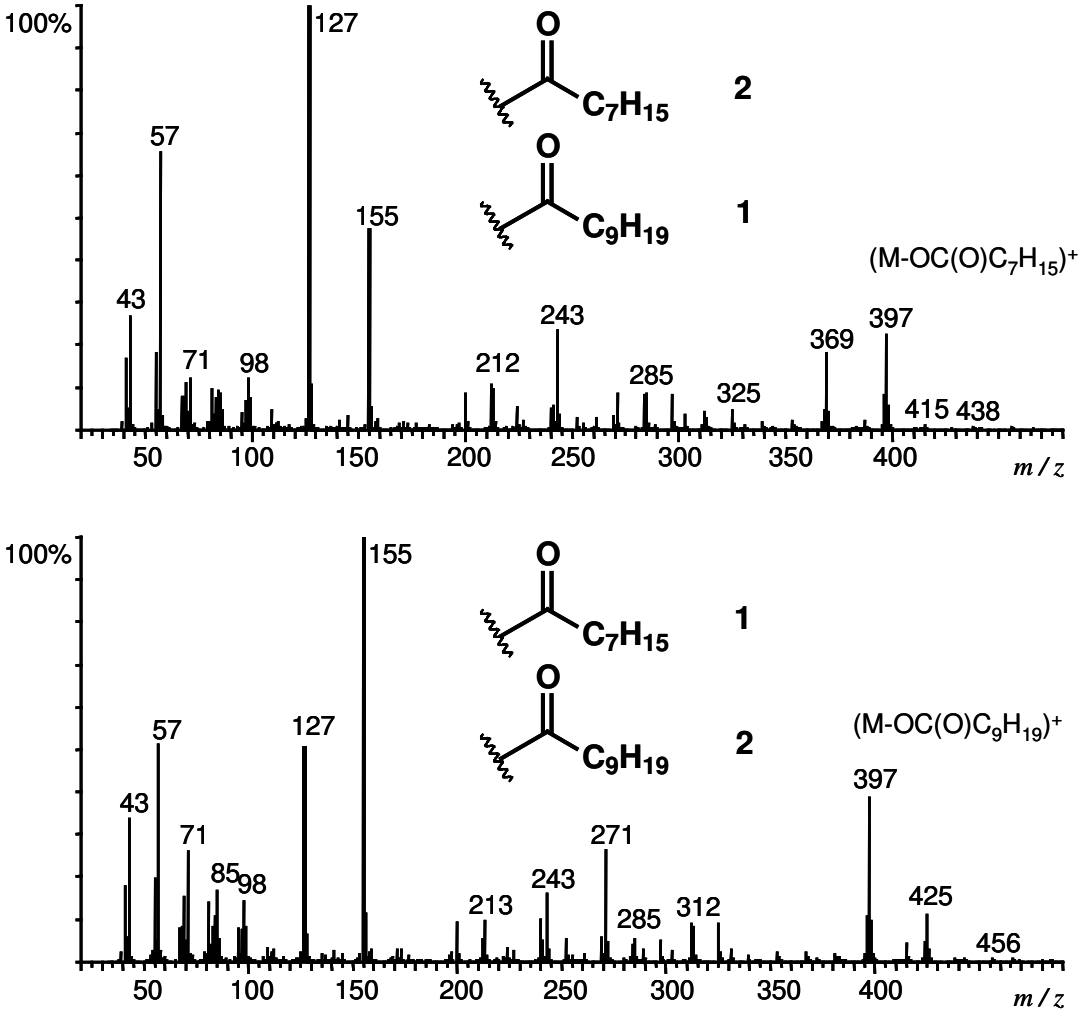


Fig. 3.16 MS spectra of **33** (M540), above, and **34** (M568), below. The intensity ratio of the acyl fragments reveals the acylation pattern of the triols (MS spectra of the other hexanetriols are in Appendix).

The esters in the sample IHg,4 were transesterified with TMSH and derivatized with MSTFA to find out which hexanetriol isomers were in the sample. The MS spectrum of the derivatization product was compared to the MS spectra of trimethylsilylated derivatives of 1,2,3-, 1,2,6- and 1,3,4-hexanetriol. None of them was the natural product.

Although the fermentation of the strains IHg,3 and IHg,4 was repeated in a large scale (5 liter) and also in different fermentation media, the production of the hexanetriol esters was not repeatable.

3.4.3.2 Long-chain alkenes and ketones from Pic009 and Bio1

Two samples, Pic009 and Bio1 (both *Cytophaga* sp.) were found to contain several isomers of long-chain ketones and related alkenes. The strains Hel73 (*Brevibacter linens*) and Pic006 (*Frigoribacter* sp.) turned out to produce one of the alkenes. The substances were identified by GC-IR, GC-MS and HRMS.

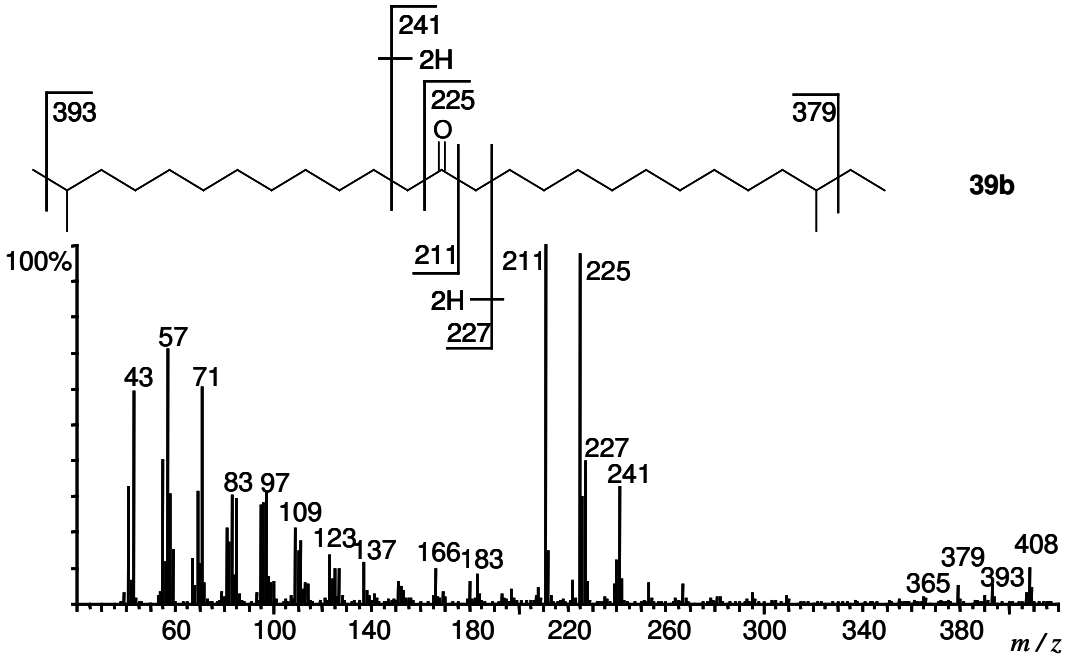


Fig. 3.17 MS spectrum of 2,24-dimethyl-14-hexacosanone (**39b**) found in Bio1.

Two or three isomers were found for each ketone with a molar mass of 366 (**36**), 380 (**37**), 394 (**38**), 408 (**39**) and 422 (**40**) and a molecular formula of $C_{25}H_{50}O$, $C_{26}H_{52}O$, $C_{27}H_{54}O$, $C_{28}H_{56}O$ and $C_{29}H_{58}O$, respectively. Characteristic ions in the MS spectra of the ketones are the α fragments (or acyl fragments) *m/z* 169, 197, 211 and 225 ($[C_{10}H_{21}CO]^+$, $[C_{12}H_{25}CO]^+$, $[C_{13}H_{27}CO]^+$ and $[C_{14}H_{29}CO]^+$) together with the corresponding fragments from the cleavage of the C_α-C_β bond accompanied by the rearrangement of two protons (*m/z* 185, 213, 227 and 241). These fragments were indicating the position of the oxo group in the chain. (Fig. 3.17)
The pattern of the alkyl fragments in the low-mass area was a sign of a longer saturated hydrocarbon chain. Nevertheless, the intensity of the alkenyl fragments, formed by proton rearrangement, indicated a branched chain. In addition, the relative high intensity of the fragments $(M-15)^+$, $(M-29)^+$ and $(M-43)^+$ showed that the branching was at the end of the chain. (Fig. 3.16)
To find out which isomers were found the retention indexes were calculated and compared to the measured ones. The retention index of a substance can be estimated according to a simple equation (Eq. 3.4):[60]

$$RI_{calc} = N + FG + \Sigma Me_i - \Sigma S \qquad \text{Eq. 3.4,}$$

where N is 100 times the length of the main chain, FG is a functional group increment, Me_i is an increment dependent on the position of the methyl branching relative to the end of the chain (Table 3.8), and ΣS is a special steric increment (5 for each 1,5 position of two methyl groups). ΣS was zero in our case.

Table 3.8 Me_i values.[60]

Position of the Me group	Me_i
2/(n-1)	60
3/(n-2)	73
4/(n-3)	56
5/(n-4)	46

The unbranched isomer, 15-nonacosanone (**41**) was isolated from cabbage by liquid extraction with hexane and the retention index was determined by GC-MS (RI = 3088). The FG value for the oxo group in 15-nonacosanone was calculated from Eq.3.4 to be 188.

The retention indexes of all the isomers of the nonacosanone were smaller than that of the unbranched 15-nonacosanone. The values measured in the sample Bio1 match very good the calculated ones for 2/(n-1), 2/(n-2) and 3/(n-2) methyl-branched 14-heptacosanones, **40a**, **40b**, and **40c**. (see Table 3.9)

Table 3.9 Measured (RI_{meas}) and calculated (RI_{calc}) retention indexes of the nonacosanone isomers found in Bio1

	RI_{calc}	RI_{meas}
15-nonacosanone (**41**)	-	3088
3,25-dimethyl-14-heptacosanone (**40c**)	3034	3031
2,25-dimethyl-14-heptacosanone (**40b**)	3021	3019
2,26-dimethyl-14-heptacosanone (**40a**)	3008	3008

The same branching pattern was found for the other ketones. Although the position of the oxo group differs slightly for different ketones (11-15), the *FG* values are still almost the same in this range, thus the retention index difference between two ketones with the same methyl branching, relative to the end of the chain, is 100 times the difference in the main chain length.

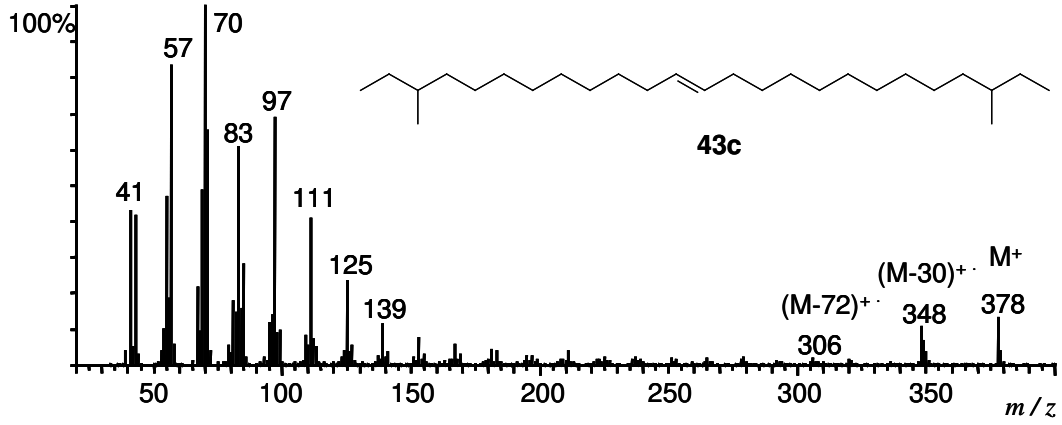


Fig. 3.17 MS spectrum of 3,23-dimethyl-12-pentacosene (**43c**: *RI* 2607) from Bio1 with the characteristic *m/z* 70 and $(M-30)^{+}$ fragments (other spectra are in Appendix).

The substances in the other isomeric series from extracts Bio1 and Pic009 had a molar mass of 364 (**42**), 378 (**43**), 392 (**44**), and 406 (**45**). Moreover, isomers of M406 were found also in the extracts of Pic006 (**45c**: *RI* 2803) and Hel73 (**45b**: *RI* 2795, **45c**: 2809). The GC-IR analysis of Pic006 (Fig. 8.28 in Appendix) showed that these substances were hydrocarbons.

The molecular weight matched the molecular formulas of an alkene series: $C_{26}H_{52}$, $C_{27}H_{54}$, $C_{28}H_{56}$ and $C_{29}H_{58}$.

Characteristic fragments in the spectra were $(M\text{-}28/30)^{+\cdot}$, $(M\text{-}44)^{+\cdot}$, $(M\text{-}56/58)^{+\cdot}$ and $(M\text{-}70/72)^{+\cdot}$ and the *m/z* 70 fragment with different intensities (Fig. 3.18, 3.19). Intensity of the fragments relative to the molecular ion was examined within a group of isomers with the same molecular mass. The relative intensity of the $(M\text{-}30)^{+\cdot}$ and $(M\text{-}72)^{+\cdot}$ ions increased while that of the $(M\text{-}44)^{+\cdot}$ and $(M\text{-}56/58)^{+\cdot}$ fragments decreased with increasing *RI* within a group. The relative intensity of the *m/z* 70 fragment increased with increasing *RI* of isomers with the same molar mass.

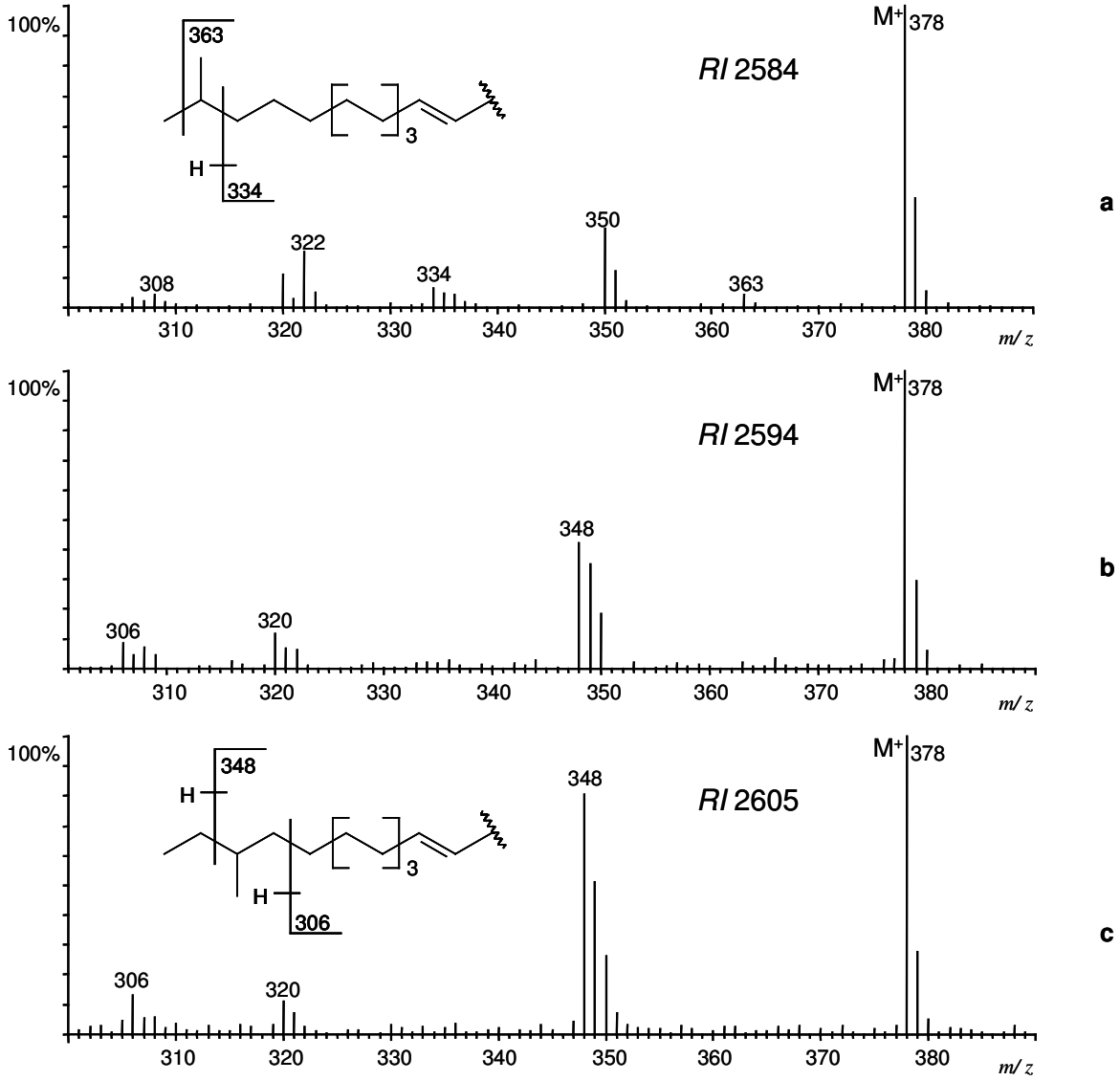


Fig. 3.19 Radical ion fragments by neutral loss from the molecular ion of the isomers of dimethyl-12-pentacosaene (**43**); **a**: 2,24-dimethyl-, **b**: 2,23-dimethyl-, **c**: 3,23-dimethyl-.

The HRMS analysis confirmed the postulated molecular formulas and also revealed the branching pattern of the alkenes. The higher intensity of the *m/z* 70 ($[C_5H_{10}]^{+\cdot}$) fragment together with the $(M-30)^{+\cdot}$ and $(M-72)^{+\cdot}$ ions indicated (n-2) methyl groups. The slightly higher intensity of the $(M-44)^{+\cdot}$ and $(M-56/58)^{+\cdot}$ fragments was a sign for (n-1) methyl groups. (See Fig. 3.19).

The position of the double bond (DB) in 3,25-dimethylheptacosene (**45c**) was determined by a micro-reaction of sample Hel73 A with iodine and dimethyldisulfide (DMDS). The generated vicinal methylthio groups were in the middle of the chain indicating position 13.

The same branching pattern and the same position of the oxo and DB functionalities were hints for a common biosynthetic origin of the ketones and the alkenes, which is discussed in more detail in chapter 3.5.

3.4.3.3 Cyclic polysulfides from Bio137 and Bio138

The extracts Bio137 and Bio138 (*Cytophaga* sp.) were found to contain cyclic polysulfides with a molar mass of 152, 208 and 240. M208 and M240 had two isomers. The molecular formula was determined by GC-HRMS analysis: $C_4H_8S_3$ (**46**), $C_8H_{16}S_3$ (**47a**, **47b**) and $C_8H_{16}S_4$, (**48a**, **48b**) each with one double bond equivalent (DBE).

The elemental composition of the fragments and the relative intensities are summarized in Table 3.10. The formula of a few fragments in parenthesis was postulated without exact mass according to literature data.[61,62] The peaks of the polysulfides were overlapping other peaks which made it difficult to get a clear MS spectrum. As a result, the relative intensity values of the fragments were averaged between the two samples. Both values are shown in the table if averaging led to high standard deviation.

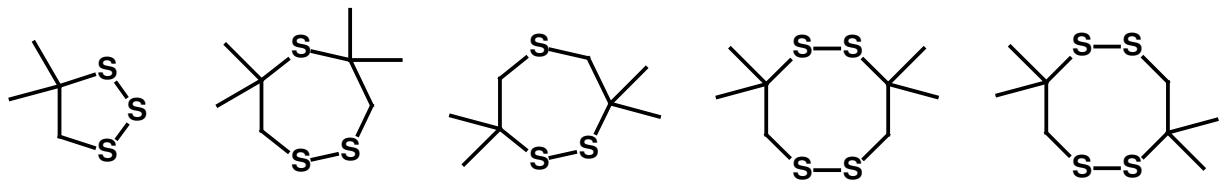
The MS spectra of the polysulfides were compared to literature data since nuclear magnetic resonance spectroscopy experiments were not possible to carry out due to the small concentration of the substances and the small amount of sample available. The MS spectra of the two isomers of M208 ($C_8H_{16}S_3$) were very similar to that of 4,4,6,6-tetramethyl-1,2,5-trithiepane (**47a**) and 3,3,7,7-tetramethyl-1,2,5-trithiepane (**47b**) described by Mieloszynski[61]. The first isomer of M240 ($C_8H_{16}S_4$) had a similar MS spectrum to that of 3,3,8,8-tetramethyl-tetrathiocan (**48a**).[61] Based on the MS data and assuming that the polysulfides found in the marine extracts are biosynthetically related a structural formula was postulated for each isomer, presented in Scheme 3.13.

Several structures were ruled out based on numerous publications containing MS data of cyclic polysulfides.[61-71]

Table 3.10 Elemental composition and relative intensity[a] of the fragments in the EI+ MS spectra of the polysulfides found in Bio137 and Bio138; the related numbered structures (Scheme 3.13) are postulated.

		M152	M208		M240	
Formula	*m/z*	**46**: *RI* 1175	**47a**: *RI* 1472	**47b**: *RI* 1516	**48a**: *RI* 1727	**48b**: *RI* 1748
$C_8H_{16}S_4^{+\cdot}$	240	-	-	-	39	62
$C_8H_{16}S_3^{+\cdot}$	208	-	100	82	-	-
$C_4H_8S_3^{+\cdot}$	152	100	71/59	100	100	100
$C_8H_{15}S^+$	143	-	13	-	-	-
$C_4H_8S_2^{+\cdot}$	120	-	28	26	24	59
$(C_3H_6S_2^{+\cdot})^2$	106	2	12	42	1	-
$S_3^{+\cdot}$	96	4	1	1	-	-
$C_4H_8S^{+\cdot}$	88	27	53	41	24	26
$C_4H_7S^+$	87	52	94	87	56	55
$C_3H_5S^+$	73	10	10	8	2	19
$S_2^{+\cdot}$	64	13	4	3	4	4
$C_2H_3S^+$	59	21	27	32	11	15
$C_4H_7^+$	55	38	95	76	54	82
$(CHS^+)^2$	45	17	19	18	9	10
$C_3H_5^+$	41	19	27	29	17	23
$(C_3H_3^+)^2$	39	18	21	20	13	16

[a] Intensity values are the average of the two measurements (Bio137 and Bio138); both values are shown if the difference is too large, for compound **48b** only data from Bio137 is shown.
[b] Formulas that were postulated without exact mass.



Scheme 3.13 Postulated structural formulas for the polysulfides found in extracts Bio137 and Bio138.

Possible 1,2,4-trithiolanes with the molar mass of 152 do not match the spectra of **46**, since the fragment $(M\text{-}C_nH_{2n-1}S)^{+\cdot}$ is missing: *m/z* 92 from the spectrum of 3,5-dimethyl-1,2,4-trithiolane[65] and *m/z* 78 from that of 3-ethyl-1,2,4-trithiolane.[68] As a consequence, and also due to the $S_3^{+\cdot}$ fragment directly formed from M^+, three adjacent sulphur atoms were postulated in the molecule of **37**. No MS data was found for the six- and seven-ring isomers (2+1) and for 4-ethyl-1,2,3-trithiolane. The vicinal five-ring isomer (3,4-dimethyl-1,2,3-trithiolane) should have two (cis/trans) isomers and thus seems less likely.

In the MS spectrum of **47a** and **47b** *m/z* 55 is the longest $C_nH_{2n-1}^{+\cdot}$ and *m/z* 87 the longest $C_nH_{2n-1}S^{+\cdot}$ fragment which means that no more than four carbons are adjacent. Considering the number of carbons (eight) in the molecule, it is only possible with two C4 units. The symmetrical 3,5-dipropyl-1,2,4-trithiolane[67,68] and the diisopropyl one[71] do not produce such a strong $(M\text{-}C_4H_8)^{+\cdot}$ fragment, *m/z* 152, but a more intensive $(M\text{-}S_2H)^{+\cdot}$ fragment, *m/z* 143. For the six-ring isomers no literature data were found. Nevertheless, the complete lack of $(M\text{-}C_nH_{2n+1})^+$ fragments near M^+ indicates aliphatic alkyl side-chain no longer than n=1, which makes the postulated seven-rings for **47a** and **47b** (Scheme 3.13) more convincing. The intensity difference of the fragment *m/z* 106 in the two isomers (Table 3.10) is also in agreement with literature data.[61] As an alternative, 3,3,6,6-tetramethyl-1,2,5-trithiepane (not presented in Scheme 3.13) can not be ruled out completely for the compound **47a** or **47b**.

The MS spectra of the polysulfides **48a** and **48b** contain the same fragments as that of **47a** and **47b** (except *m/z* 106), which is a hint for a similar structure. Since the spectrum of **48a** matches the literature data of 3,3,8,8-tetramethyl-tetrathiocan,[61] 3,3,7,7-tetramethyl-tetrathiocan is a realistic suggestion for **48b**.

3.4.3.4 Isomeric lactones from AMP,3 and Hel59

Two isomeric substances **49a** and **49b** (*RI* 1454 and 1468) were found with a molar mass of 209 and almost identical MS spectrum in the extracts of the strains AMP,3 (*Cytophaga* sp.) and in Hel59 (*Marinomonas* sp.).

An intensive carbonyl band was detected for **49a** and **49b** by GC-IR at 1801 cm^{-1} and 1804 cm^{-1}, respectively, which probably indicates an unsaturated γ-lactone structure.[72] (Fig. 8.29 and 8.30 in Appendix)

The molecular formula and the elemental composition of the fragments (see Table 3.11) was determined by GC-HRMS analysis: the M^+ is $C_{12}H_{19}NO_2$ with four or five DBE.

The concentration of the compounds was higher in the polar fraction of the extracts. Besides, two additional isomers **50a** and **50b** (*RI* 1442, 1462) were detected in Hel59, the MS spectra were similar to that of **49a** and **49b**. (The related MS spectra are in Appendix.)

Table 3.11 Elemental composition and relative intensity of the fragments in the EI+ MS spectra of the substances **49a** and **49b** in AMP,3 analyzed by HRMS. -: no HRMS data available.

Formula	*m/z*	Relative intensities	
		49a	**49b**
$C_{12}H_{19}NO_2^{+\cdot}$	209	7.7	10.3
$C_{11}H_{16}NO_2^{+}$	194	0.8	0.6
$C_{11}H_{19}NO^{+\cdot}$	181	3.6	5.7
$C_9H_{12}NO_2^{+}$	166	60.0	45.3
$C_8H_{10}NO_2^{+}$	152	4.0	6.0
$C_8H_{12}NO^{+}$	138	12.8	9.0
$C_6H_6NO_2^{+}$	124	67.8	52.8
-	110	3.9	3.6
$C_6H_{10}O^{+\cdot}$	98	31.8	47.7
$C_5H_7O^{+}$	83	11.6	13.0
$C_4H_5O^{+}$	69	24.9	31.2
$C_3H_3O^{+}$	55	100	100
-	43	24.9	22.5
-	41	30.8	27.5

3.5 Discussion

3.5.1 Data analysis tool

Hundreds of lipophilic bacterial extracts were derivatized and analyzed by GC under standardized conditions. The program ChromConv converted the GC data in report format into a measurement matrix. The converted GC data were processed by SPSS and the similarity of the GC patterns was visualized by dendograms created by HCA and PCA/HCA

methods (section 3.4.1). No *a priori* knowledge was at hand about the nature of the data set. Not even the fermentation conditions were given, since marine bacteria are extremely difficult to culture and thus many different media were used. The problem presented in this work was therefore more elaborate than other described chromatographic applications of clustering (section 3.2.2). It belongs to the field of exploratory data analysis, and interpretation of the results should be taken with caution.

Data selection

Identification of as many components as possible was carried out by GC-MS analysis of selected samples in order to obtain information about the metabolism of marine bacteria. Several extracts were chosen for further investigation based on properties, like unusual smell, such as Pic009, or biological activity, for example Bio137. Other samples were picked out according to their position in the dendogram and analyzed by GC-MS in order to seek novel substances. The extract of the strain ANT17 (subset STM3), for example, contained mostly alkanes, alkenes and fatty acids. On the other hand, Bio1 (subset STM3) was found to produce the same methyl-branched long-chain alkenes and ketones as Pic009 (presented in section 4.3.2.2)

Once an extract was found to contain novel substances, similar ones were sought by clustering. Nevertheless, the same interesting compounds (or series of compounds) were often found in extracts with different overall GC patterns. Thus clustering a particular part of the GC chromatograms was necessary to find other samples containing the given compounds, as in case of the two C35-terpenes **26** and **27** (Dendogram 8.12 made by analyzing the 84-89 min segment of the chromatograms).

Interpretation of the results

It is difficult to measure validity by using unsupervised clustering techniques. The methods measure rather the goodness of the clustering,[24b,25b] that is how good the results reveal the natural grouping in the data set, which was in our case secondary. Since none of the components of the marine extracts were preferred (most of them were unknown), it was not possible to prepare model mixtures and perform cluster analysis on a model data set. Several strains were repeatedly fermented, some extracts were repeatedly derivatized, and a few were repeatedly injected, which helped to evaluate our data selection tool.

Similar but somewhat irregular GC patterns, as in case of the extracts IHg,3, IHg,4 and ACF,5 (see Fig. 3.10), were found to be similar by the method, just like most of the repeatedly derivatized and injected samples. Yet, not all of the derivatization and injection control samples were clustered together (see Bio210 control samples in Dendogram 8.7 and 8.8), which means the method is relatively sensitive to concentration differences and impurities from the vials and the GC system (see Fig.3.20).

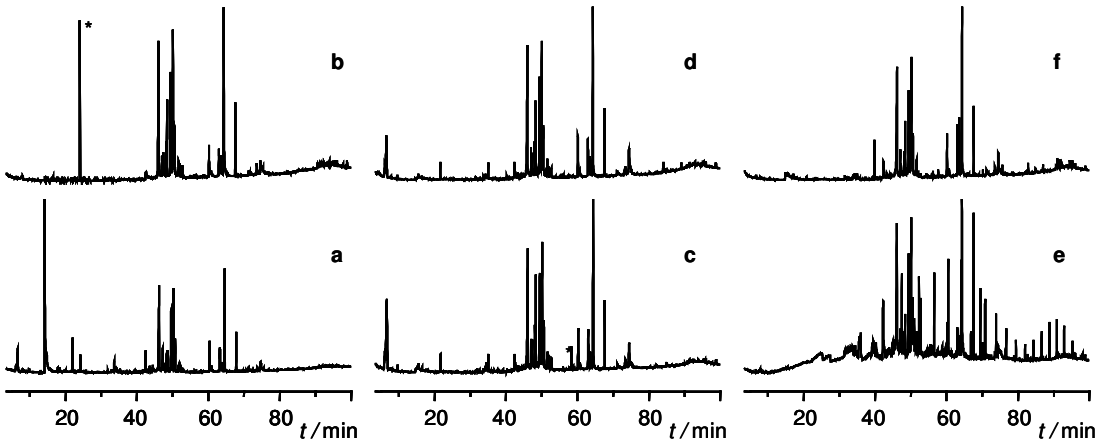


Fig. 3.20 GC patterns of Bio210 (**a**), Bio210 der.I (**b**), Bio210 der.IIa/1 (**c**), Bio210 der.IIa/2 (**d**), Bio210 der.IIb/1 (**e**), Bio210 der.IIb/2 (**f**). Spikes from detector are marked by *. Bio210 was derivatized repeatedly (**a**, **b**, **c**, **e**); **c** injected once more is: **d**; **e** injected once more is: **f**. It can be seen that early peaks are not reproducible and by the injection of **e** impurities could be observed.

The GC pattern of the strains fermented repeatedly in the project was rarely reproducible (see Fig.3.21 and sample names in bold in Dendogram 8.11 in Appendix). It is well-known that some bacteria do not produce identical compounds under identical culture conditions for unknown reasons.

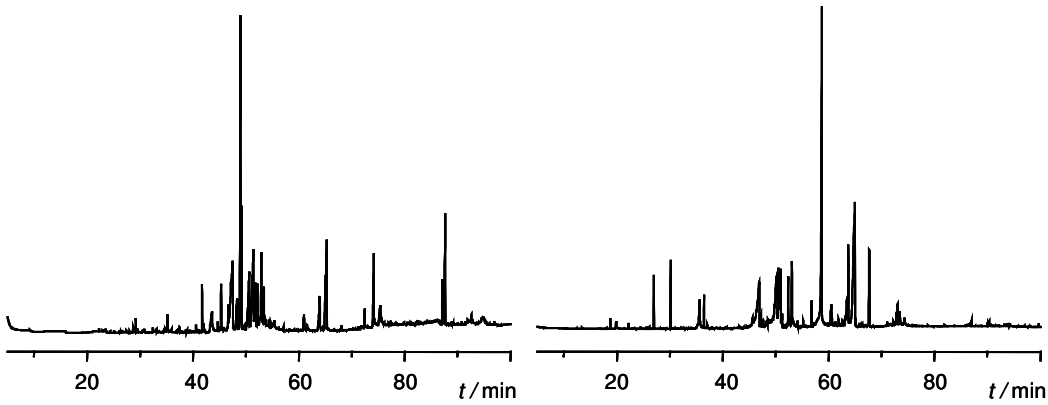
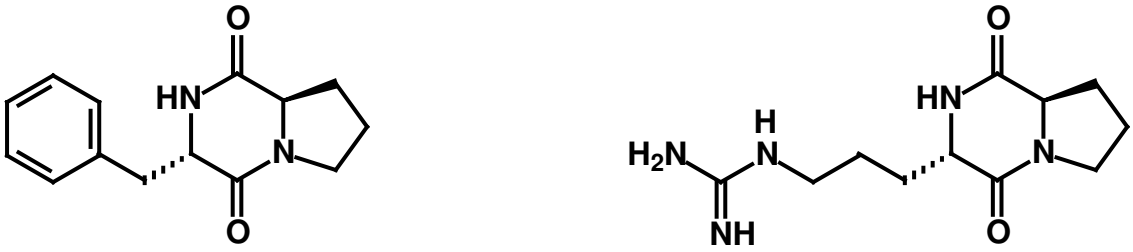


Fig. 3.21 GC patterns of the repeatedly fermented strain Hel45 (*Roseobacter* sp.).

High background in the GC analysis

Limitations of the clustering tool arise mainly from the nature of the fermentation experiments. The culturing conditions, the composition of the medium and the solvent used for extracting the fermentation mixture play an important role in the reproducibility of the experiment and the successfulness of data selection by cluster analysis.

The origin of the problems was the high background in GC analysis. Large number of components unspecific to the strains were evenly distributed in the middle part of the GC chromatograms, roughly in the 30-80 min segment of the total interval. Typical substance classes were fatty acid derivates, diketopiperazines, sugar derivates (less in the lipophilic fractions), and aromatic substances (presented in table 3.3 in section 3.4.1.1). Diketopiperazines are well-known natural products biosynthesized generally from proteinogenic L-α-amino acids by cyclization of the corresponding L,L-dipeptide.[73] As exceptions to the rule, cyc(L-Xxx-D-Pro) diketopiperazines have been also isolated (see Scheme 3.14), but the origin of them is still unclear. Since proline racemases are rarely found, a non-enzymatic pathway is more plausible.



51 **52**

Scheme 3.14 Unusual cyc(L-Xxx-D-Pro) diketopiperazines as natural products: **51**, cyc(L-Phe-D-Pro), isolated from the fungus *Aspergillus flavipes*, and **52**, cyc(L-Arg-D-Pro), produced by the bacteria *Pseudomonas sp*. IZ208.[73]

Moreover, formation of the thermodynamically more stable cyc(L-Xxx-D-Pro) from the cyc-L,L-epimer under mild acidic or basic conditions was confirmed.[74-77]

The diketopiperazines found also in control samples without bacteria in our experiment (see Table 3.x in section 3.4.1.1) gave always two peaks in the GC analysis. By GC-MS analysis of reference compounds the occurrence of the epimers cyc(L-Leu-L-Pro) and cyc(L-Leu-D-Pro) was demonstrated, with lower concentration of the second. Hence, it is most unlikely, although not impossible that the diketopiperazines identified in the marine extracts were of bacterial origin. These substances are more polar and thus give a broad chromatographic peak

(with leading). They were the most disturbing components due to overlapping other, probably more interesting peaks.

In order to minimize the influence of the background substances on the results the time segments containing peaks of diketopiperazines and artifacts (phthalates and aromatic phosphates) were excluded from the data mining process by ChromConv for HCA. Nevertheless, omitting all the segments with disturbing substances would have resulted in loosing substantial chemical information. The effect of the diketopiperazine peaks on the clustering is demonstrated in the dendograms of subset STM3 and STM5 (Dendogram 8.6 and 8.10 in Appendix) obtained by the PCA/HCA method: extracts with high diketopiperazine concentration were clustered together, since here the whole time interval was included in the clustering process.

3.5.2 Conclusion

The lipophilic bacterial extracts investigated in the marine research project in Lower-Saxony contained various novel substances and a high background originated from fermentation media. To apply unsupervised learning techniques for the exploratory analysis of such a complex data set a better experiment design is necessary.

It is very important to control the repeatability of the fermentation process, of the derivatization method and the GC analysis. The most critical point is to which extent a strain produces the same substances under identical conditions. Parallel experiments should be carried out with a few strains under the same conditions and also over a period of time (day-to-day reproducibility) and the results should be evaluated by the data mining tool.

The GC pattern was often dominated by the substances from the fermentation medium. Therefore it is advisable to note the fermentation parameters and perform cluster analysis with extracts obtained from fermentation experiments with the same medium composition. It is also required to make fermentation blanks (without bacteria) to each medium on a regular basis to obtain information about the background. Cluster analysis should be carried out including these blank extracts.

Adding a proper internal standard to the mixture at the beginning of the fermentation experiment helps to control the yield. It provides then an internal standard for sample preparation as well. Deuterated alkanes or fatty acids would be suitable, for example.

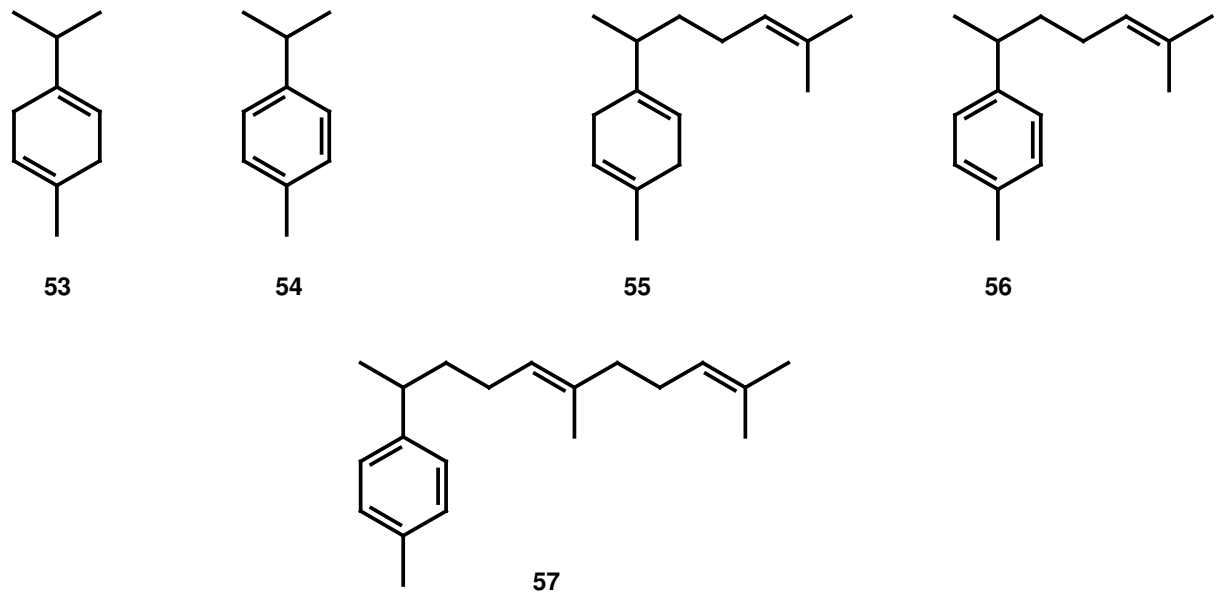
Selecting a short- and a long-chain isomer would cover the range of volatility and the GC peaks would not interfere with others.

3.5.3 Interesting substance classes from marine bacteria

Several interesting compounds were found and identified during the marine research project (described in sections 3.1 and 3.4.3). By analyzing the lipophilic fractions of the fermentation mixture in the second period of the project two terpenes, **26** and **27**, were isolated and their structure was elucidated. Other compounds are not yet fully investigated, still it is interesting to discuss their possible origin and biosynthesis.

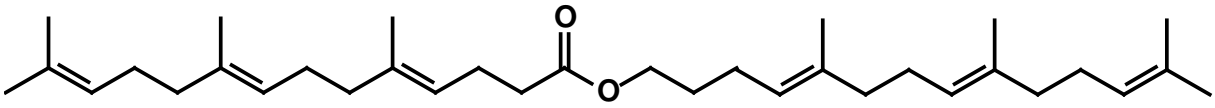
C35 terpenes

Well-known short-chain homologues of **26** and **27** are the monoterpenes γ-terpinene (**53**) and p-cymene (**54**) from essential oils of *Citrus*, *Origanum*, and *Thymus* species,[78,79] the sesquiterpenes β-curcumene (**55**)[80] and *ar*-curcumene (**56**)[80,81] found in various essential oils of *Curcuma* and *Salvia* species,[43] and the diterpene **57** isolated from *Salvia dorisiana* (Scheme 3.15).[82]

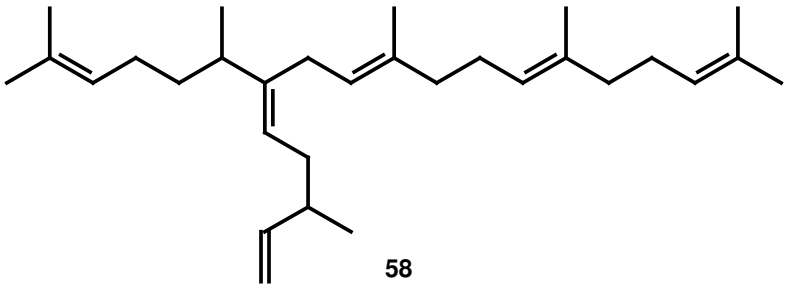


Scheme 3.15 Homologues of the C35 Terpenes **26** and **27** : γ-terpinene (**53**), p-cymene (**54**), β-curcumene (**55**), *ar*-curcumene (**56**), (*E*)-1-methyl-4-(1,5,9-trimethyl)-deca-4,8-dienylbenzene (**57**).

Long-chain isoprenoids are often found in marine sediments, they probably regulate membrane fluidity of the cell.[83] The so called highly-branched isoprenoids (HBI) with different chain-length (C-20, -25, -30) and saturation are biomarkers characteristic to the source organism and the different environmental conditions.[84-86] An example is the compound **58** isolated from the marine diatom *Rzizosolenia setigera* (Scheme 3.16).[87] Isoprenoid wax esters, such as compound **59**, are known from *Acinetobacter*, *Pseudomonas* and *Marinobacter* species (Scheme 3.16).[88]



59



58

Scheme 3.16 An isoprenoid wax ester, **59**, from *Marinobacter squalenivorans*, and a C-30 HBI, **58**, from the marine diatom *Rzizosolenia setigera*.

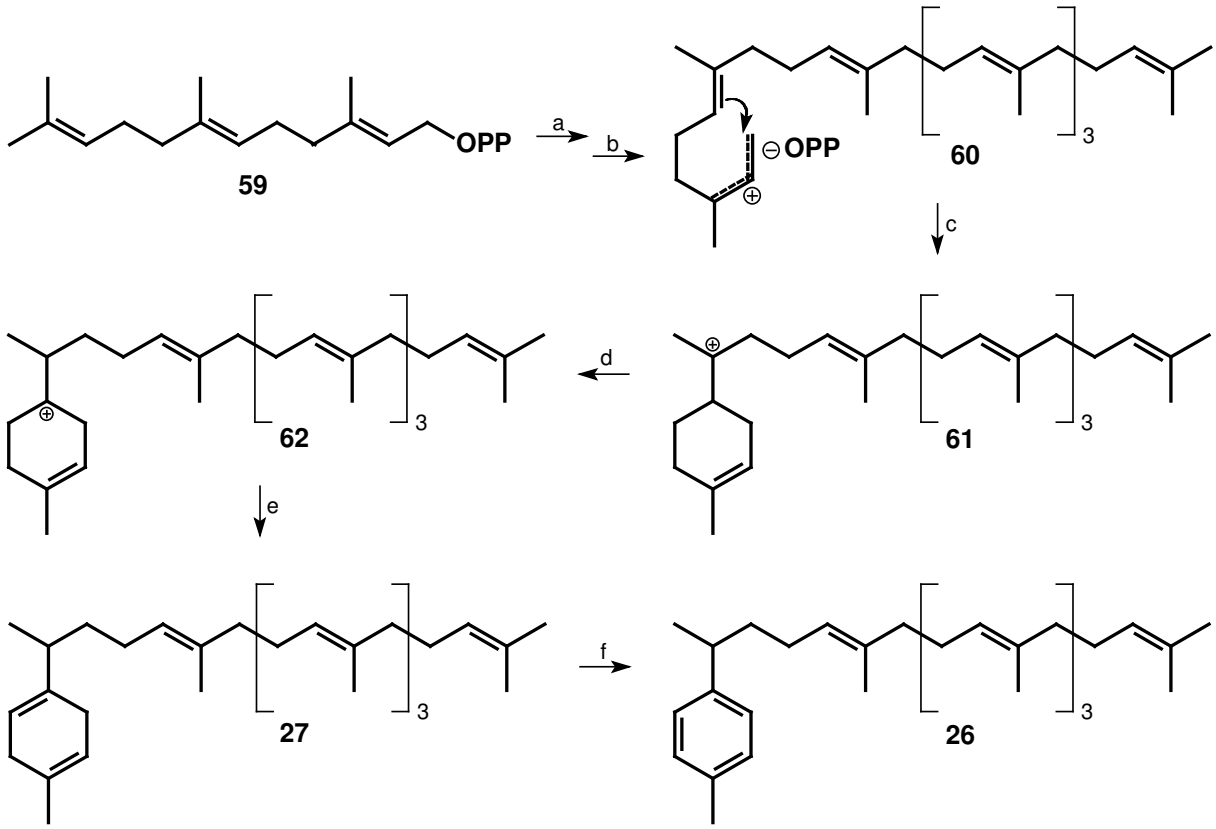
The biosynthesis of longer (>C-20) isoprenes is quite versatile. Aliphatic C35 isoprene hydrocarbons are biosynthesized by head-to-head coupling of farnesyl pyrophosphate (FPP) and geranylgeranyl pyrophosphate (GGPP) by the deep see methanogen *Methanococcus jannaschii* under anaerobic conditions.[83] It was also observed that a C-30 carotenoid synthase accepts longer substrates, GGPP for example, and is thus able to produce a C35 backbone.

Nevertheless, aliphatic head-to-tail all-*trans*-C-30 and C35 isoprene chains are produced by elongation of FPP with (three or four) isopentenyl pyrophosphate (IPP) units. The reaction is catalyzed by prenyl-transferase type II enzymes (consisting of two components) discovered in *Bacillus subtilis*.[89a]

Moreover we assume that the final cyclization and aromatization steps of the biosynthesis of **26** and **27** are mechanistically similar to the biosynthesis of γ-terpinene and p-cymol from geranyl pyrophosphate (GPP).[89b]

According to the literature data a biosynthetic pathway for **26** and **27** was postulated. (Scheme 3.17) From FPP (**59**) (generated by an FPP-synthase) an all-*trans* heptaprenyl chain is

produced by stepwise elongation with IPP mediated by an extra prenyl-transferase (**a**). A terpene cyclase regulates the conformation of the chain and mediates the following reactions: isomerization of the carbocation **60** (**b**), nucleophilic attack of the double-bond and production of carbocation **61** (**c**), generation of carbocation **62** from **61** by 1,2-hydrid shift (**d**), and formation of the product **27** by deprotonation (**e**). The hexadiene **27** is eventually aromatized to **26** by a desaturase (**f**).



Scheme 3.17 Postulated biosynthetic Pathway to **26** and **27**. Step **a**: FPP (**59**) is elongated by four IPP units by prenyl transferase II. Steps **b**-**e** are catalyzed probably by a terpene cyclase; **b**: isomerization of the carbocation and regulating the conformation of the chain, **c**: cyclization, **d**: 1,2-hydride shift, and **e**: deprotonation. Step **f**: the terpene **27** is aromatized to **26** by a desaturase or non-enzymaticaly by oxygen.

According to our observation the concentration of the terpene **27** was higher than that of the aromatic derivative **26** in freshly fermented samples, but the relation has turned to the opposite over time. For this reason it is still uncertain whether **26** is a natural product or a more stable oxidation product of **27**.

Long-chain ketones and alkanes

The occurrence of the structurally related long-chain ketones and alkenes in the extract of the strain Bio1 (summarized in Table 3.12.) reflects their biosynthetic origin (Scheme 3.18). It is well-known that (n-1)- and (n-2)-methyl-branched fatty acids (FAs) are synthesized from special starter units generated from isobutyryl-CoA, 2-methylbutyryl-CoA and 3-methylbutyryl-CoA, which are catabolic products of the amino acids L-Val, L-Ile and L-Leu respectively.[90] As a consequence of the nature of the chain-elongation (always by a C-2 unit), the (n-2)-methyl-FAs have an even-numbered backbone starting from 2-methylbutyryl-CoA. The (n-1)-methyl-branching occurs both by even- and odd-numbered backbones due to the corresponding even and odd starting units (3-methylbutyryl-CoA and isobutyryl-CoA). As a result, a few isomers are not produced by the bacteria.

Table 3.12 Summary of the ketones and alkenes found in Bio1 (*Cytophaga* sp.) with the measured *RI* values; -: not detected.

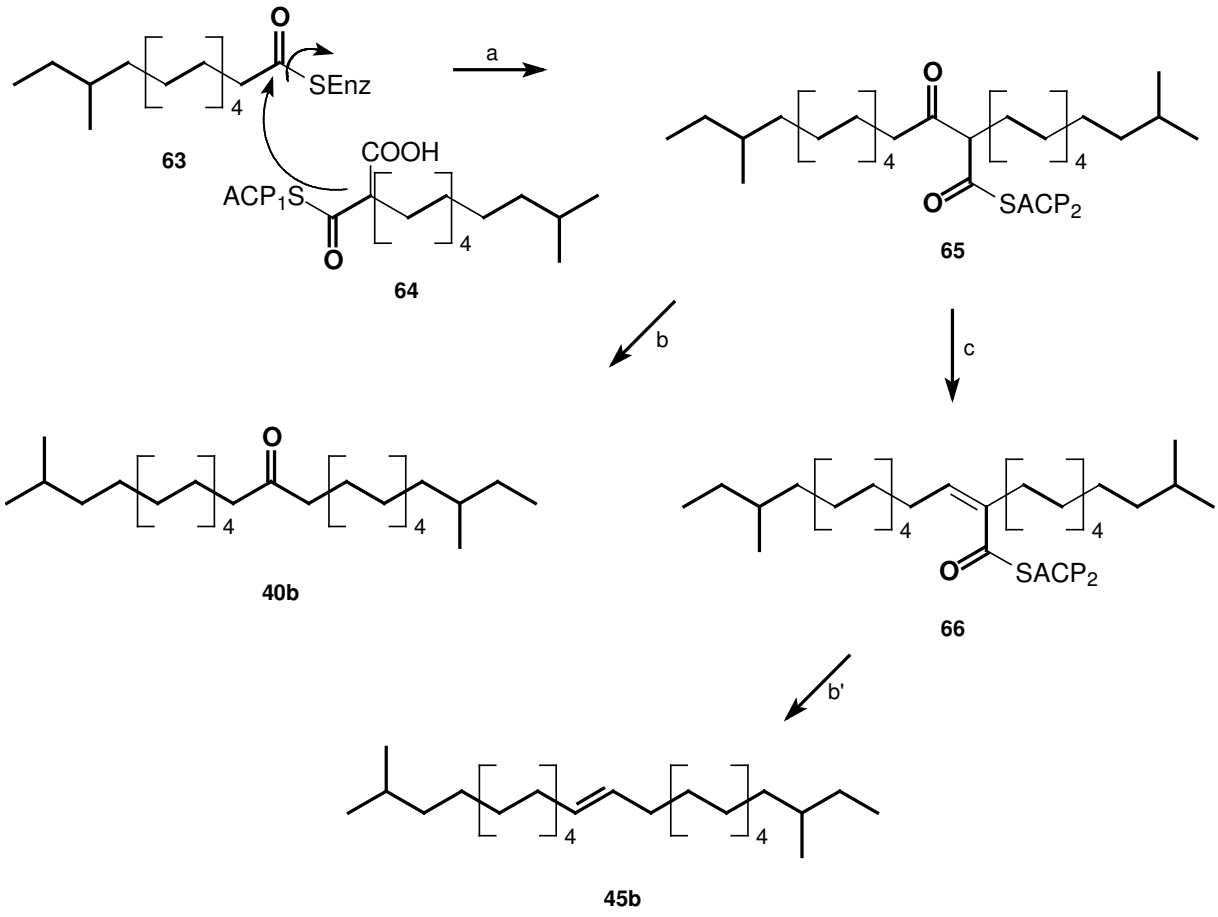
Position of the Me groups	Chain-length	Ketones	RI_{meas}	Alkenes	RI_{meas}
3, (n-2)	27	3,25-dimethyl-14-heptacosanone **40c**	3031	3,25-dimethyl-13-heptacosaene **45c**	2807
2, (n-2)		2,25-dimethyl-14-heptacosanone **40b**	3019	2,25-dimethyl-13-heptacosaene **45b**	2795
2, (n-1)		2,26-dimethyl-14-heptacosanone **40a**	3008	2,26-dimethyl-13-heptacosaene **45a**	2784
3, (n-2)	26	3,24-dimethyl-13-hexacosanone[c]	-	3,24-dimethyl-13-hexacosaene[d]	-
2, (n-2)[a]		2,24-dimethyl-13-hexacosanone **39b**	2920	2,24-dimethyl-13-hexacosaene[d] **44b**	2696
2, (n-1)		2,25-dimethyl-13-hexacosanone **39a**	2908	2,25-dimethyl-13-hexacosaene[d] **44a**	2686
3, (n-2)	25	3,23-dimethyl-12-pentacosanone **38c**	2828	3,23-dimethyl-12-pentacosaene[d] **45c**	2607
2/3, (n-2/1)[b]		2,23-dimethyl-12-pentacosanone **38b**	2817	2,23-dimethyl-12-pentacosaene[d] **43b**	2595
2, (n-1)		2,24-dimethyl-13-pentacosanone **38a**	2807	2,24-dimethyl-12-pentacosaene[d] **43a**	2585
3, (n-2)	24	3,22-dimethyl-12-tetracosanone[c]	-	3,22-dimethyl-12-tetracosaene[d]	-
3, (n-1)[a]		3,23-dimethyl-12-tetracosanone **37b**	2717	3,23-dimethyl-12-tetracosaene[d] **42b**	2495
2, (n-1)		2,23-dimethyl-12-tetracosanone **37a**	2707	2,23-dimethyl-12-tetracosaene[d] **42a**	2485
3, (n-2)	23	3,21-dimethyl-10-tricosanone **36c**	2625	-	-
2/3, (n-2/1)[b]		2,21-dimethyl-10-tricosanone **36b**	2615	-	-
2, (n-1)		2,22-dimethyl-10-tricosanone	-	-	-

[a] only one isomer is possible
[b] both dimethyl isomers, 2,(n-2) and 3,(n-1), are possible
[c] isomer probably not produced
[d] postulated double-bond position

The postulated mechanism of the biosynthesis of 2,25-dimethyl-14-heptacosanone (**40b**) and 2,25-dimethyl-13-heptacosaene (**45b**) is shown below (Scheme 3.18). Both building blocks are derived from FAs commonly found in bacteria. In essence, two long-chain FAs are

condensed head-to-head in a fashion similar to Claisen-condensation. This reaction has so far not reported from nature. 13-Methylmyristic acid is activated by addition of a carboxylic acid group to form an alkylmalonate, as known for acetyl-CoA in normal FA biosynthesis.[90] This compound (**64**), bound to an acyl-carrier protein (ACP), is then condensed with 12-methylmyristyl-thioester to form the keto ester **65**. Via decarbonylation the ketone **40b** is released. By reduction and dehydration of **65**, well-known reactions from fatty acid synthases (FASs), the α,β-unsaturated **66** is produced and the alkene **45b** is released by decarbonylation.

The mechanism is being currently investigated vie feeding experiments, due to the fact that one of the strains, Hel73, produces 3,25-dimethyl-13-heptacosaene (**45c**) reproducibly.



Scheme 3.18 Postulated biosynthetic route to the ketone **40b** and the alkene **45b**. (**a**) Condensation reaction of 12-methylmyristyl-thioester (**63**) and activated 13-methylmyristyl-SACP (**64**), the β-keto-adduct **65** is generated. (**b**) Decarbonylation of **65** to the ketone **40b**. (**c**) Reduction and dehydration of **65**, the α,β-unsaturated **66** is produced. The alkene **45b** is released by decarbonylation (**b'**) of **66**.

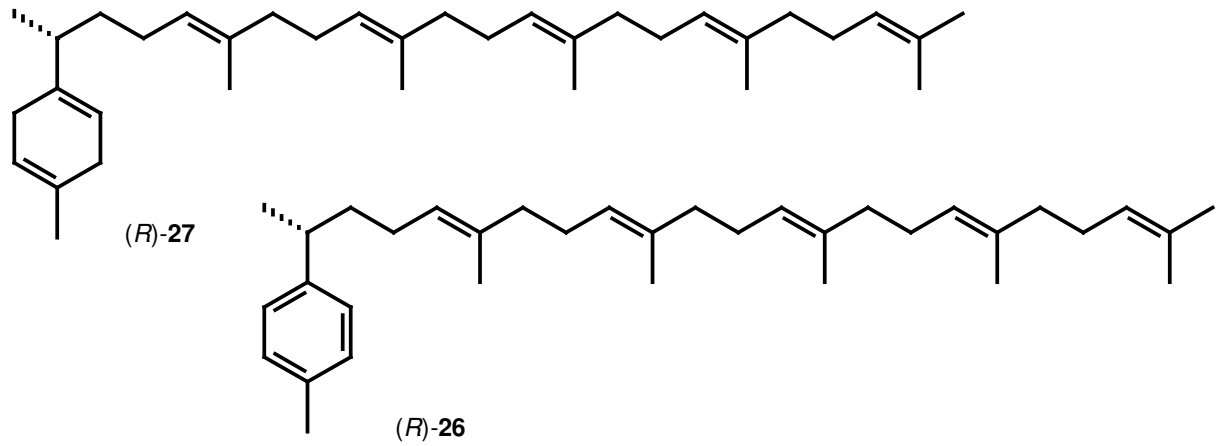
3.6 Summary

More than 500 bacterial extracts were analyzed by means of unsupervised learning techniques based on their gas chromatographic pattern. Raw GC data was arranged into a measurement matrix by a self-made program ChromConv. Hierarchical cluster analysis was successfully applied on converted GC data for selecting interesting samples for further analysis (GC-MS, GC-IR, NMR) and finding similar ones to save measurement labor and time.

Although a few clusters of samples were revealed during clustering (the hexanetriols in Fig. 3.10), it was not the clusters we were looking for. Hierarchical cluster analysis was rather used as a presentation method of the similarity degree between bacterial fermentation extracts according to their GC pattern.

Two methods were used. Variables were extracted manually by ChromConv or by PCA of the original variables (by using SPSS). Hierarchical cluster analysis was performed by SPSS on the whole data set as well as on smaller subsets with samples fermented more or less at the same time (see dendograms in Appendix). Despite the high background in the GC measurements originating from the fermentation medium several interesting substance classes were discovered.

Two C35 terpenes, (*R*)-tetraprenyl-β-curcumene ((*R*)-**27**) and (*R*)-tetraprenyl-*ar*-curcumene ((*R*)-**26**) (Scheme 3.19) were isolated by liquid chromatographic methods and their structure was elucidated by GC-MS and NMR analysis.



Scheme 3.19 (*R*)-tetraprenyl-*β*-curcumene ((*R*)-**27**) and (*R*)-tetraprenyl-*ar*-curcumene ((*R*)-**26**).

According to the cluster analysis results on the relevant segment of the GC chromatograms several extracts were found to contain these substances (Dendogram 8.12 in Appendix), which helped the isolation. The absolute configuration was determined by chiral GC analysis of the degradation product of **27**, we assume that **26** has the same configuration: *R*. Monocyclic hydrocarbon terpenes with this chain-length have been hitherto unknown. The strains producing these substances possess probably a unique terpene cyclase enzyme.

A series of long-chain ketones and biosynthetically related alkenes with (n-1) and/or (n-2) methyl branching at both chain-ends were identified (Table 3.12 in Discussion). A special condensing enzyme and a FAS supposedly mediate the unusual biosynthetic pathway. (Scheme 3.18) Although the production of these substances by the strains Bio1 and Pic009 do not seem to be reproducible, other strains are found to synthesize the alkene **45c**. The postulated mechanism is being currently studied.

Other substance classes have not been yet fully investigated. The tentative structures of the hexanetriols (Scheme 3.12 in section 3.4.3.1) and the cyclic polysulfides (Scheme 3.13 in section 3.4.3.3) must be synthetically confirmed, since isolation was not possible due to the low amount of extract available and the lack of reproducible production of these compounds.

4 LOW MOLECULAR WEIGHT COMPONENTS IN THE VENOM OF THE SPIDER *CUPIENNIUS SALEI*

4.1 Spider venom

Introduction

Spiders, species from the order Aranea, are one of the most widespread living creatures on Earth; they are carnivores and produce offensive material, venom, to capture their prey.[91] Although from the more than 30 000 species described so far just about 180 bite man and only the minority is truly venomous,[91,92] they are treated with respect by humans (from caution to arachnophobia). Dangerous is the bite of widow spiders (genus *Latrodectus*), Australian funnel-web spiders (genus *Atrax*), banana spiders (genus *Phonentria*) and members of the genus *Loxosceles*.[91,92] Since spiders do not attack primarily men and due to the good supply of antiserum, severe injuries are more frequent from stings of honey-bees in Australia, the habitat of the majority of these threatening species.

Meanwhile, scientists are interested in finding the basis of the success of these animals in evolution. Although investigations have started already in the middle of the 20th century, the examination of spider toxins gained considerable attention in the last decades.91 It is very tedious to obtain suitable amount of venom for analysis (the yield is sometimes less than 1 µl pro animal) due to the small size of spiders and their glands. Nevertheless, as a result of the fast improvement of analytical techniques, and the development of modern, nondestructive milking methods,[93,94] substantial knowledge has been gained about the composition and biological role of spider venom since the 1980's.

Neuroactive compounds of different substance classes (proteins, polyacylamines) have been found in the venom of several spider species. These substances have reversible or irreversible paralyzing effect on prey animals, thus they are good candidates for insecticides. Moreover, a bite of the most dangerous species may cause death to humans when not properly treated. Still, purpose of the research is not solely to seek for antivenomous compounds, to generate novel insecticides or to understand the behavior of spiders but to find useful tools to investigate neural diseases, such as stroke, Alzheimer's disease, and epilepsy.[16]

Composition of spider venom

The amount of venom available from one animal varies from a few hundred nanoliters to several microliters. It is common that females produce more venom than males, also when body size is taken into consideration.[95] Venom composition is highly variable. Interspecific, intersex and interpopulational differences can be significant not only due to different venom constituents but also to different concentration ratios.[95]

Spider venom contains generally proteins with or without enzymatic activity, polypeptides, low molecular weight compounds (<1000 Da) (such as purine bases, nucleotides, amino acids, citric acid) and inorganic salts.[16,91,96]

The peptide content of the venom has been more thoroughly investigated, compared to the group of low molecular weight (LMW) compounds. Among the well investigated neurotoxic acting proteins some are selective blockers of presynaptic ion channels, others form divalent cation channels.

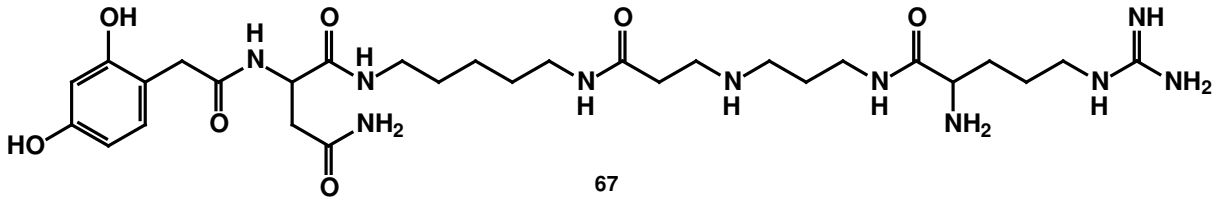
Several examples are mentioned here to illustrate different modes of action. Both the agatoxin ω-Aga-IA (7.5 kDa, 66aa) from the funnel web spider *Agenelopsis aperta* and the peptide SNX-325 (49 aa) from *Segestria florentina* block N-type Ca ion channels, and the former one inhibits also L-type Ca channels.[91] Proteins from *Latrodectus* and *Steatoda* species bind to presynaptic membranes and form transmembrane ion channels, which leads to liberation of neurotransmitters.[97] Two toxins from the black widow spider *Latrodectus mactans* act similarly but with different prey selectivity: α-latrotoxin (130 kDa) is active on vertebrates, α-latroinsectotoxin (120 kDa) on insects.[91] Other peptides, such as μ-Aga-I (4.2 kDa) (*A. aperta*) and two peptides (PhTx2-5, PhTx2-6, 6-8 kDa) from the American banana spider *Phonentria nigriventer* act on voltage-sensitive Na ion channels and cause repetitive release of transmitters from nerve cells. [91,97]

In addition to neuroactive peptides a few enzymatic active proteins have been found, such as hyaluronidases,[98] phosphplipases, proteases, peptide isomerases, and collagenases.[91]

Typical LMW substance classes found in spider venom are amines, polyamines and free amino acids. Some of the LMW compounds contribute to synergistic effects. Histamine, for example, increases the permeability of capillary walls and allows toxin transport, whereas taurine inhibits neurotransmitters in the prey animal, decreasing the chance of escape.[98,99] Citric acid is also a known constituent of spider venom, it prevents self-intoxication by inhibition of phospholipases.[91]

A relatively new group of LMW compounds, the acylpolyamines were discovered in the beginning of the 1980s. Acylpolyamines have been isolated mostly from the venom of spiders

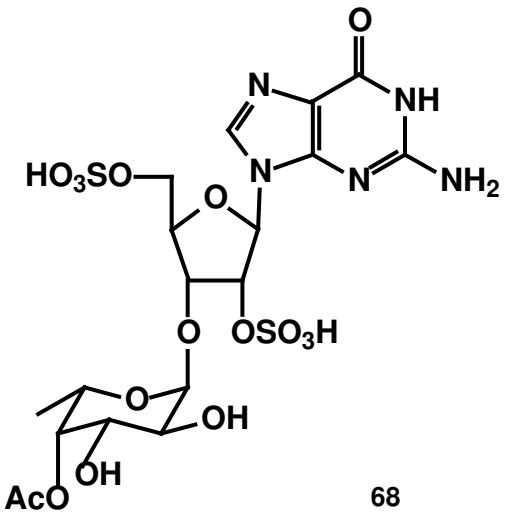
in the family of Araneidae and Agelenidae.[97] The two main structural constituents of these compounds are a lipophilic aromatic acyl head group and a polyamine chain. Additional parts are optional: an amino acid linker between the head group and the chain, and a basic amino acid or positively charged amino group at the other end of the chain, such as in compound **67**. (Scheme 4.1).[93] Acylpolyamines are neurotoxic and bind selectively to neural receptors or ion channels.[97] Most of them block glutamate receptors and thus inhibit Ca ion channels.[91]



67

Scheme 4.1 The acylpolyamide Argiotoxin 636 (**67**) found in the poison gland of *Argiope lobata*.[97]

In 1993, an unusual neuroactive compound, HF-7 (**68**), was isolated from the venom of a funnel-web spider, *Hololena curta*. This compound is a glyconucleoside disulfate, namely the 3'-O-(4''-O-acetyl-α-L-fucopyranosyl)-guanosine-2',5'-disulfate (Scheme 4.2), which also blocks glutamate-sensitive Ca channels.[100,101]



68

Sheme 4.2 HF-7 (**68**) from *Hololena Curta*.

Certain neural disorders (Alzheimer's disease, for example) are connected to the misbehavior of glutamate receptors, therefore these neuroactive components of spider venom might be useful for further investigation of such diseases.[97]

Analysis of low-molecular weight components of spider venom

For the isolation and analysis of LMW venom constituents a mixture of chromatographic and spectroscopic methods is used. The peptide content is usually separated from the rest by gel filtration, for example, and several steps of purification, such as cation exchange chromatography, RP-HPLC, TLC, CE are applied to isolate the compounds.[98]

Identification of LMW compounds in venom fractions is possible by HPLC-MS or by derivatization and GC-MS, although the latter is not suitable for analyzing thermally labile compounds, acylpolyamines for instance. For structure determination tandem mass spectrometry (MS-MS) is required, since the soft ionization techniques used in HPLC-MS do not result in substantial fragmentation. As an example, in case of the acylpolyamines isolated by RP-HPLC, the position of nitrogen atoms in the chain was resolved by continuous-flow fast atom bombardment MS-MS (FAB MS-MS) analysis.[91,93]

The exact structure is determined usually by NMR analysis of the natural compound, and it is eventually confirmed by synthesis.

It is important to avoid loosing venom during isolation due to the low amount available. Using micro-HPLC for the isolation of venom constituents or analyzing the crude venom directly can be a solution to the problem. The molecular mass can be measured by matrix assisted laser desorption ionisation mass spectrometry (MALDI-MS)(SP011), and structural information can be obtained by low- and/or high energy collision induced dissociation (CID) even without HPLC separation.[93] Recent coupled techniques allow to obtain amazing amount of structural information from one injection. Online HPLC-UV(DAD)-API-MS/MS systems are good alternatives to the ’old’ FAB techniques.[16,17]

Snake venom and venom proteins from spiders have been investigated by HPLC,[102,103] also by CE.[104,105] However, the use of CE for the examination of LMW components of spider toxins has not been reported so far. This technique is especially promising, since it is suitable for analyzing polar metabolites in low amounts mostly without time consuming sample preparation.

4.2 Introduction in capillary electrophoresis

Basics

High-performance capillary electrophoresis (HPCE) has been probably the most rapidly developing analytical separation technique in the last two decades. It covers a group of methods based on different migration velocities of charged particles. The movement of the different entities is driven by the electric field applied between the two ends of a capillary, which is filled with a buffer solution, also called the background electrolite.

Separation occurs nowadays in an uncoated or chemically modified fused silica capillary with typical internal diameters of 50-100 μm. A basic case of introducing electric force (0-30 kV) in such a narrow tube filled with a BGE and analytes are demonstrated in Fig. 4.1.

The inner capillary wall is usually negatively charged because of dissociated silanol groups on the inner surface with the exception of very acidic conditions or due to additives. It attracts positive ions from the solution and an electric double-layer forms. When applying voltage the diffusion layer start to move to the negative electrode, the cathode. Then, owing to viscosity, the whole bulk solution moves in the same direction, to the cathode, with a constant velocity after reaching a stable voltage in a few seconds. It is called the electroosmotic flow (EOF) and its mobility (and velocity) has a positive sign in the case described above. It is also possible to eliminate or reverse the electroosmotic flow by changing the capillary wall: by lowering the pH of the BGE, adding additives (dynamic coating) or by chemical modification of the silica surface.[106a,107a]

Neutral molecules are taken by the electroosmotic flow, whereas charged analytes migrate in the electrical field to the anode or the cathode according to their ionic mobility and the EOF. The mobility of cations is positive, of anions negative. The net electrophoretic velocity of a charged analyte in the EOF is the sum of the individual electrophoretic velocity and the velocity of the electroosmotic flow.

Analytes are separated according to their charge and size under given conditions such as capillary type, composition of the BGE, applied voltage and temperature.

A special feature of the electrically driven flow is its flow profile. Whereas liquid chromatographic techniques generate parabolic flow, the liquid current in CE is constant in the axial direction, except a rapid flow drop to zero at the capillary wall.[107a] Therefore very high theoretical plate numbers are possible to achieve in HPCE: in the order of 10^5–10^6. Nevertheless, extra heat is produced owing to molecular collisions in the high field strength

applied in the electrolite. This phenomenon is called the Joule heating. To support heat dissipation and thus eliminate the resulting peak broadening, the capillary has to be cooled.

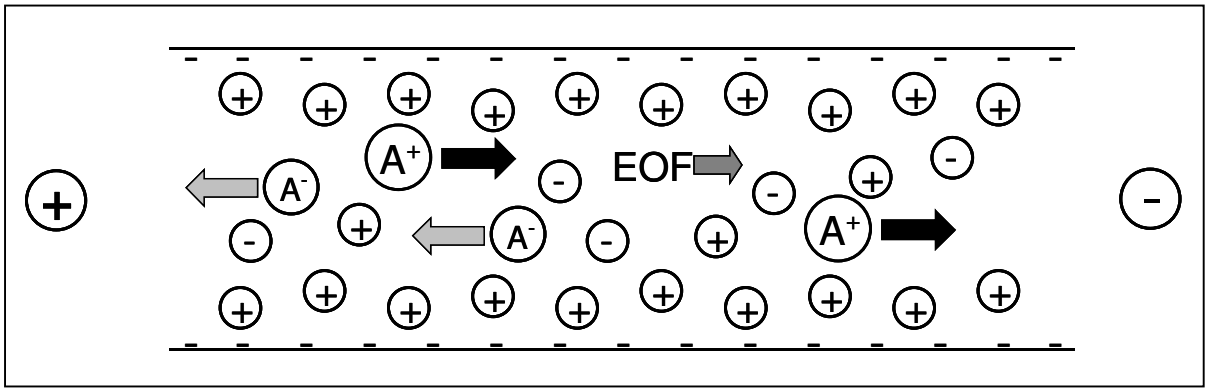


Fig. 4.1 Direction of the velocities of analytes A^+, A^- and the electroosmotic flow (EOF) in a bare fused silica capillary after applying positive voltage.

High efficiency, easy quantitation and automation make HPCE superior to slab gel electrophoresis, although 2D protein separations still remain a challenge.[106b]

HPCE has many advantages over HPLC, too. It is usually faster, more efficient, needs lower amounts of solvent and analyte. Analysis of biological samples is very convenient due to aqueous buffers mostly used as BGE. On the other hand, migration time of the analytes is not always reproducible since it is sensitive to chemical changes of the capillary wall. In addition, scaling-up separations to preparative level is easier by HPLC.[107b]

Although the concentration limit of detection of CE is worse (higher) than that of HPLC, CE is much more mass sensitive.[107c] Thus CE-MS is a very promising technique, although not yet routinely used.

Instrumentation

Current HPCE systems are fully automated. Samples and buffer vials are stored on large plates and the measurement-cycle (Fig. 4.2) is carried out by the autosampler. First the rest of BGE from previous measurement is removed and the capillary wall is conditioned by rinsing with basic or acidic solutions by applying pressure or vacuum. As next, the capillary is filled with the BGE followed by injection of the samples by applying low pressure (hydrodynamic injection) or electric field (electrokinetic injection). Then separation occurs by merging both ends of the capillary into the vials containing BGE solution filled to the same level.

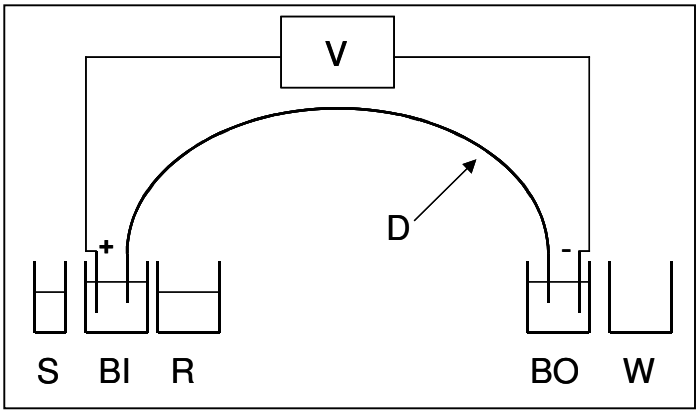


Fig. 4.2 A typical measurement cycle in HPCE. S: sample vial, BI and BO: buffer inlet and outlet vials, R: rinse vial, W: waste vial, V: high-voltage power supply, D: detector. The inlet end of the capillary is by the positive electrode, the anode. The steps are after washing the capillary: 1. filling the capillary with BGE from R to W with pressure, 2. injection of the analyte solution from S to W with pressure or vacuum, 3. separation from BI to BO by applying voltage (positive in this figure).

The temperature of the capillary is maintained constant by air-cooling, or even better, a coolant fluid is circulated in the cartridge holding the capillary. In some instruments the samples can be cooled, too. A high-voltage supply provides the longitudinal electric field in the capillary for the separation. Ultraviolet absorbance, photo-diode-array, fluorescent, conductivity and mass spectrometry detectors are the commercial ones in capillary electrophoresis systems.[107d] The detection window is usually a few centimeters from the separation end, except for conductivity and MS detection.

Different techniques and applications

Capillary zone electrophoresis (CZE), micellar electrokinetic chromatography (MEKC), capillary gel electrophoresis (CGE), capillary isoelectric focusing (CIEF), capillary isotachophoresis (CITF) and capillary electrochromatography (CEC) are all capillary separation techniques, where the bulk flow is electrically driven.[107b] The main characteristics of these methods are presented in Table 4.1. They are also possible to perform in the same system in most cases.

Due to its high efficiency and various modes of action, high-performance capillary electrophoresis competes already with high-performance liquid chromatography in several application fields, such as clinical analysis, peptide mapping, analysis of LMW compounds using detergents and by MEKC, chiral separations, carbohydrate analysis, and separation of inorganic ions.[106b]

Table 4.1 Different modes of HPCE.[107e]

	Principle	Application
CZE	separation according to charge to size ratio of the analytes	small molecules in general, peptides, proteins, carbohydrates, enantiomer separations[107f]
MEKC	uncharged analytes migrate in micelles owing to surfactant concentrations over cmc	universal except DNA, enantiomer separations[107g]
CGE	macromolecular analytes are separated according to their size in the gel- or polymer network-filled capillary	proteins, oligo- and polynucleotides, complex carbohydrates[107h]
CIEF	analytes separate according to their isoelectric point in a longitudinal pH gradient in the capillary	peptides, proteins, smaller zwitterionic substances[107i]
CITP	analyte zones are moving between a leading and a terminating zone with the same velocity, dilute analytes are enriched	mostly a trace enrichment technique during injection in CZE[107j]
CEC	uncharged molecules are separated on a stationary phase	universal except DNA, enantiomer separations[107k]

4.3 Investigation of the venom of *Cupiennius salei* by different separation techniques

Cupiennius salei

One of the best investigated spiders today is the wandering spider *Cupiennius salei* (Araneae:Ctenidae) (Fig. 4.3) found originally in the tropical rainforest from Central Mexico to northern South America.[94,98] It came to Europe with banana transports.

The yield of electrical milking from adult females is 7-15 µl venom per animal. It has been shown that the spider is able to use its venom economically according to the size and mobility of its prey.[94]

The concentration of the inorganic ions Na and Ca in the venom of *C. salei* is lower, that of the K ion is higher than in the hemolymph of spiders in general. This reveals a possible change in the function of the toxins after being injected into the hemolymph of prey animals.[98]

Fig. 4.3 *Cupiennius salei* consuming its prey.[108]

LMW compounds identified so far in the venom of *C. salei* are free amino acids (Gly, Asx, Ala, Arg, Leu, for example), amines, such as histamine and taurine, and the polyamines putrescine and cadaverine in very low amounts.[98]

The neuroactive peptides isolated form the venom of *C. salei* can be divided into two groups: some peptides (6-8 kDa) act neurotoxic, others, the so called cupiennins (3-4 kDa), have an additional cytolitic activity.[94] The most toxic component is CSTX-1 (8.4 kDa), a highly basic peptide (pI 9.26) containing 74 amino acids and four disulphide bridges.(SP015) It has an inhibitory effect on L-type Ca ion channels. The cupiennins are characterized by the lack of Cys, and most of them lack also Pro. These peptides are active against Gram + and Gram – bacteria and show cytolitic activity against human erythrocytes as well. They bind to the membrane with their polar C terminus and the insertion of the hydrophobic end (N-terminus) destroys the membrane structure.[94] Although cupiennin1a is less neurotoxic than CSTX-1, the two peptides act probably synergistically.

One larger protein (with a molecular mass 33-38 kDa) was found to have hyaluronidase activity in enzyme assay experiments.[98] This protein is most likely a spreading factor in target tissue.

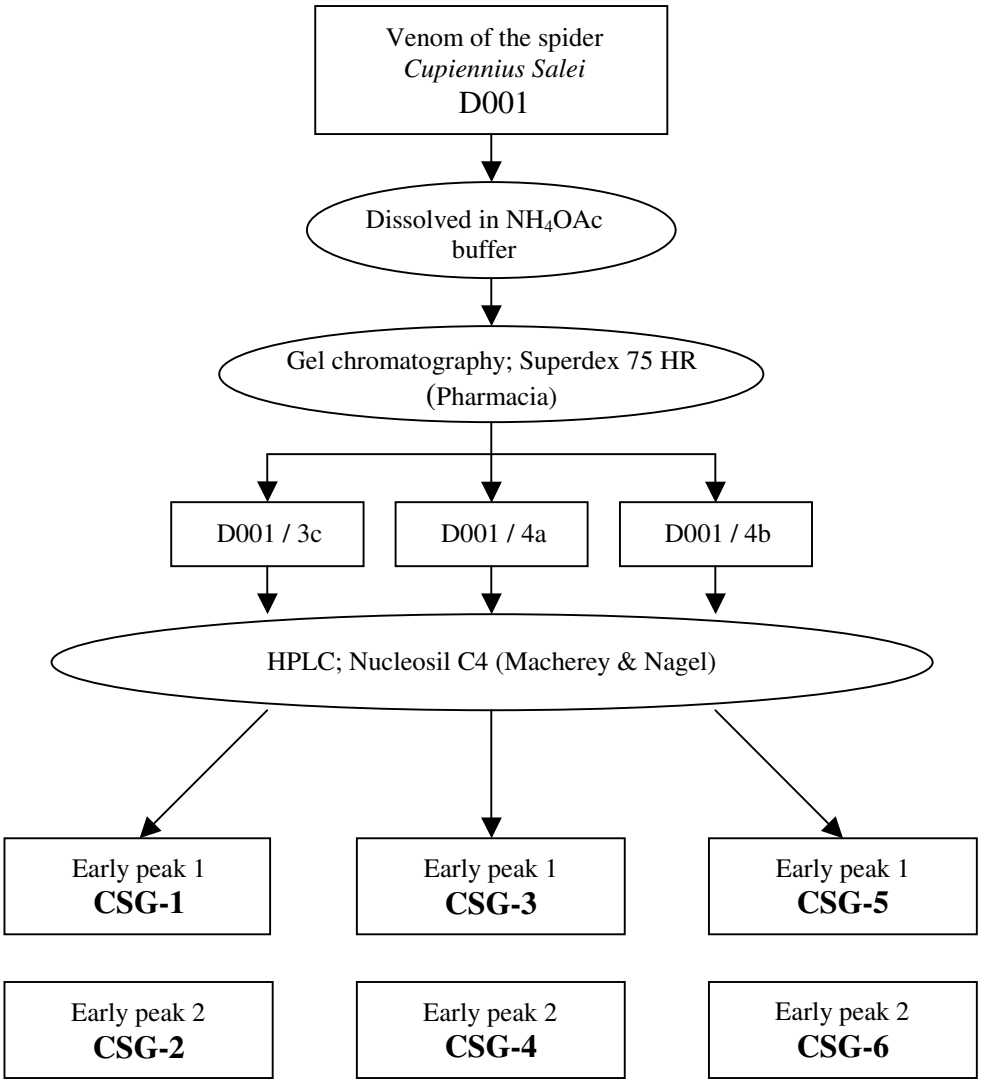
Isolation of LMW fractions of *C. salei* by the group of Prof. Nentwig

Venom composition of *C. salei* has been thoroughly investigated by the research group of Prof. Nentwig from the Institute of Zoology, Department of Synecology, University of Bern

(Switzerland).[98,94] In order to examine the LMW content of the venom more closely, specific fractions and a crude venom sample were sent to our laboratory.

The venom fractions containing LMW components were gained by several chromatographic separation steps from the crude venom of *C. salei*. (Scheme 4.3) Spider venom was dissolved in an ammonium acetate buffer and the components were separated according to their size on a Superdex 75 HR 10/30 (Pharmacia) column. Three middle fractions (3c, 4a, 4b) were chosen for further purification on a Nucleosil 300-5C4 column (4mm x 250mm, Macherey & Nagel). Molecular weight of the retained compounds were between 2-4 kDa.

Since are aim was to identify LMW components, the early eluting two peaks were collected for our research purposes. The fractions were washed more times with distilled water and then lyophilized. In our experiments we refer to these samples as CSG-1 to CSG-6. Analysis of the fractions by GC-MS and CE-UV(DAD) was planned and samples were kept at $^{-}80^{o}$C until then.



Scheme 4.3 Isolation of the low molecular weight fractions CSG-1 to CSG-6 from the venom of the spider *Cupiennius salei* by the group of Prof. Nentwig.

4.3.1 GC-MS experiments

Venom fractions

The samples CSG-1 to CSG-6 were derivatized with MSTFA without solvent and analyzed by GC-FID and GC-MS (Table 4.2). Two pairs of fractions were similar; CSG-1 and CSG-3 contained citric acid, and CSG-4 and CSG-6 contained histamine as main component. In the samples CSG-2 and CSG-5 no substance was detected by GC-MS.

In order to verify the identification of the compounds reference substances were used in control experiments; commercially available citric acid, isocitric acid and histamine were derivatized with MSTFA.

Table 4.2 Substances found in derivatized CSG-fractions and reference samples analyzed by GC-MS. TMS: trimethylsilyl, TFA: trifluoracetyl derivative. For +/- signs see Table 3.3.

	CSG-1 + MSTFA	CSG-2 + MSTFA	**CSG-3** + MSTFA	**CSG-4** + MSTFA	CSG-5 + MSTFA	**CSG-6** + MSTFA	c-Aconitic acid + MSTFA	t-Aconitic acid + MSTFA	Citric acid + MSTFA	Isocitric acid + MSTFA
Phosphoric acid, tri-TMS-	++	-	+	-	-	-	++	++	++	++
3-Methylene-succinic anhydrid	+	-	-	-	-	-	+	++	+	+
Lactic acid, di-TMS-	+	-	+	-	-	++	-	-	-	-
Succinic acid, di-TMS-	-	-	+	-	-	-	-	-	-	+
2-Methylene-succinic acid, di-TMS-	++	-	+	-	-	-	+++	+++	+++	+
Isocitric acid lactone, di-TMS-[a]	-	-	-	-	-	-	-	-	-	+++
c/t-Aconitic acid, tri-TMS-	+	-	+	-	-	-	+++	+++	+	++
Unknown acid, x-TMS-	+++	-	+	-	-	-	-	-	+	-
Citric acid, tetra-TMS-	+++	-	+++	-	-	-	-	-	+++	-
Isocitric acid, tetra-TMS-	+	-	+	-	-	-	-	-	-	+++
Histamine	-	-	-	++	-	-	-	-	-	-
Histamine, *N*-TFA-	-	-	-	+++	-	+++	-	-	-	-
Glycerine, tri-TMS-	-	-	+	+	-	-	-	-	-	-
Myristic acid, mono-TMS-	-	-	+	-	-	-	-	-	-	-
Pentadecanoic acid, mono-TMS-	-	-	+	-	-	-	-	-	-	-
Palmitic acid, mono-TMS-	(+)	-	+	-	-	-	-	-	-	++
Heptadecanoic acid, mono-TMS-	-	-	(+)	-	-	-	-	-	-	-
Oleic acid, mono-TMS-	-	-	+	-	-	-	-	-	-	++
Stearic acid, mono-TMS-	-	-	(+)	-	-	-	-	-	-	+

[a] Two peaks of diastereomers.

Although citric acid in CSG-1 and CSG-3, and histamine in CSG-4 and CSG-6 were confirmed, the existence of other short chain acids in fractions CSG-1 and CSG-3 remained uncertain. The compounds 2-methylene-succinic acid, its anhydride, succinic acid, and aconitic acid were identified in CSG-1 and CSG-3, and also detected in the derivatized reference samples of citric and isocitric acid. Hence it was not obvious whether these compounds are really constituents of the venom of *C. salei*, or are formed during derivatization or chromatographic separation.
The geometric isomers of aconitic acid were not separated on the BPX-5 phase.

Crude venom

The crude venom sample was analyzed by GC-MS. Three aliquots of about 2 µl from the sample were transferred into 1 ml vials and derivatized with MSTFA, diazomethane solution in dichloromethane, and TMSH solution in methanol.
Derivatization with diazomethane was successful, nevertheless no other substance as citric acid was detected. The results of the other derivatization experiments were not significant.

4.3.2 Capillary electrophoresis experiments

The venom fractions CSG-1 to CSG-6 were further analyzed by capillary electrophoresis to confirm GC-MS results. Two separate methods were used: the one for negative inorganic ions and small organic acids, the other for positive inorganic ions and amines. Indirect UV detection was required to analyze compounds without a chromophore functionality. Samples were spiked with magnesium chloride in order to have an internal standard to follow fluctuations of migration time.

4.3.2.1 Literature

For the analysis of small organic acids and inorganic anions indirect detection is usually used. Typical chromophores with suitable electrophoretic mobility are aromatic acids, such as salycic acid,[109] benzoic acid,[110] phthalic acid,[111] trimellitic acid,[112] and pyromellitic acid.[113,114]
The electroosmotic flow is lowered or reversed to the anode by adding modifiers to the BGE

solution; such additives are tetradecyltrimethylammonium bromide,[109,112] myristyltrimethylammonium bromide,[111] and hexamethonium hydroxide.[113,114]

Inorganic cations and amines are frequently analyzed together by CE with indirect UV detection. Chromophores in case of smaller amines are, for example, copper(II)-sulfate[115] and imidazol,[116] whereas the BGE for the indirect analysis of biogenic amines contains quinine sulfate.[117,118]

It is possible to analyze anionic and cationic analytes simultaneously if certain criteria are met; the absolute value of the mobility of the electroosmotic flow must be slightly smaller or larger than that of the anionic analytes, and two chromophores are required to provide indirect UV detection.[119]

4.3.2.2 Conditions of the CE experiments

Negative ions

A commercially available buffer from Fluka was used to analyze samples in the negative ion mode at a pH value of 7.7. The content of the buffer was the following:

- pyromellitic acid 2.25 mM,
- hexamethonium hydroxide 0.75 mM,
- triethanolamine 1.6 mM,
- sodium hydroxide 6.5 mM.

The electrophoretic mobility of the buffer (electroosmotic flow) was measured with methanol as neutral analyte. Once it was injected at the usual end of the capillary and 30kV was applied for separation, followed by injection at the short end, 10 cm from detector window, and separation at $^{-}$30kV. The measured values were $28 \cdot 10^{-5}$ $cm^2(Vs)^{-1}$ and $23 \cdot 10^{-5}$ $cm^2(Vs)^{-1}$, respectively. Notwithstanding the difference between the two values, it showed an electroosmotic flow in the direction of the cathode. Since the cathode (negatively charged) was at the injection end, the negative analytes were measured in counter-flow by applying negative voltage.

Positive ions

To measure positive ions and amines we used an other commercially available buffer from Fluka. The pH value of the buffer was set to 3.0 with phosphate. Buffer constituents were:

- copper(II) sulfate 4.0 mM
-18-crown-6-ether 3.0 mM,
- formic acid 4.0 mM.

Indirect detection at 217 nm made the analysis of both organic and inorganic cations possible. Electroosmotic flow was measured at ⁻20 and 20 kV from the short and long end, and turned out to be $13 \cdot 10^{-5}$ $cm^2/(Vs)$ and $16 \cdot 10^{-5}$ $cm^2/(Vs)$, respectively. It was a low electroosmotic flow, again to the cathode, the bulk flow and the cationic analytes migrated in the same direction by applying positive voltage.

4.3.2.3 Results

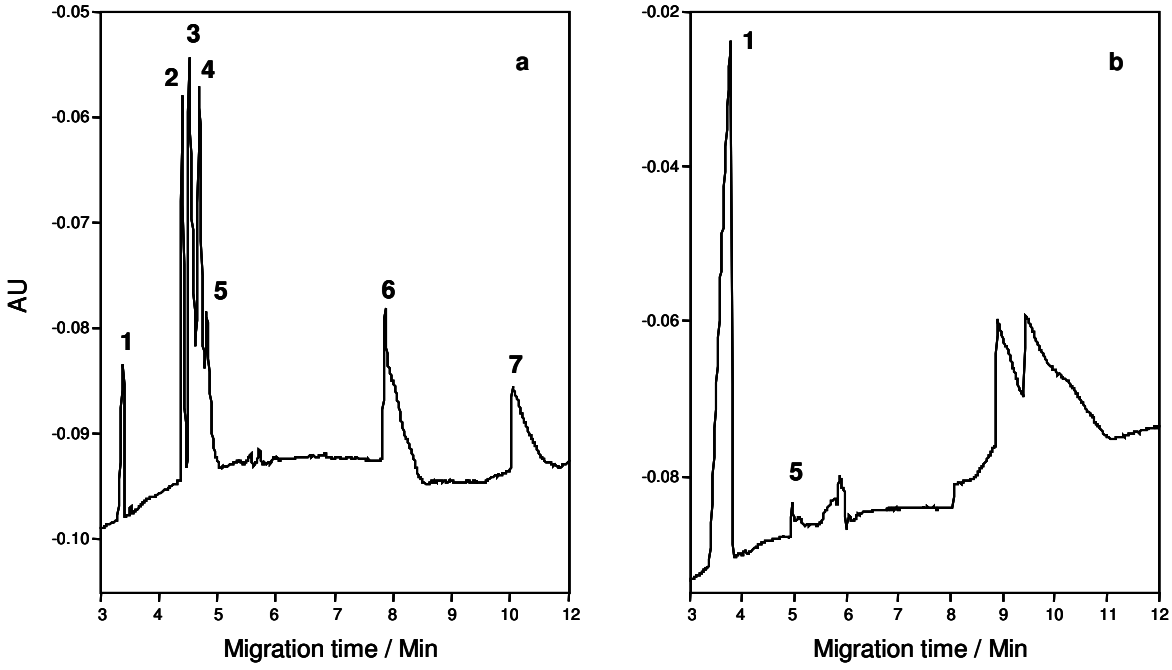


Fig. 4.4 Separation of the anion standard (**a**) and the venom fraction CSG-3, spiked with Cl^- (**b**) in negative ion mode. Components: (**1**) Cl^-, (**2**) *t*-aconitate, (**3**) *c*-aconitate, (**4**) isocitrate, (**5**) citrate, (**6**) phosphate, and (**7**) acetate. Injection: 0.5 psi, 5s; separation: Fluka anion buffer, ⁻30kV; detection: UV indirect at 220 nm.

Negative ions

The separation of the four anions *t*-aconitate, *c*-aconitate, isocitrate, and citrate, identified in fractions CSG-1 and CSG-3 by GC-MS, was acceptable with the original Fluka buffer (Fig.

4.4a) Adding Ca^{2+} in a concentration of 0.1 mM to the buffer resulted in a better separation (Fig. 4.5b), but the decrease in reproducibility of the migration times of late eluting ions, citrate, phosphate, and acetate, became significant. None of the isomers aconitic acid was found in the venom fractions. Citrate was detected as the only acid anion in CSG-1 (Fig. 4.5a,b), and in small amounts in CSG-3 (Fig. 4.4b).

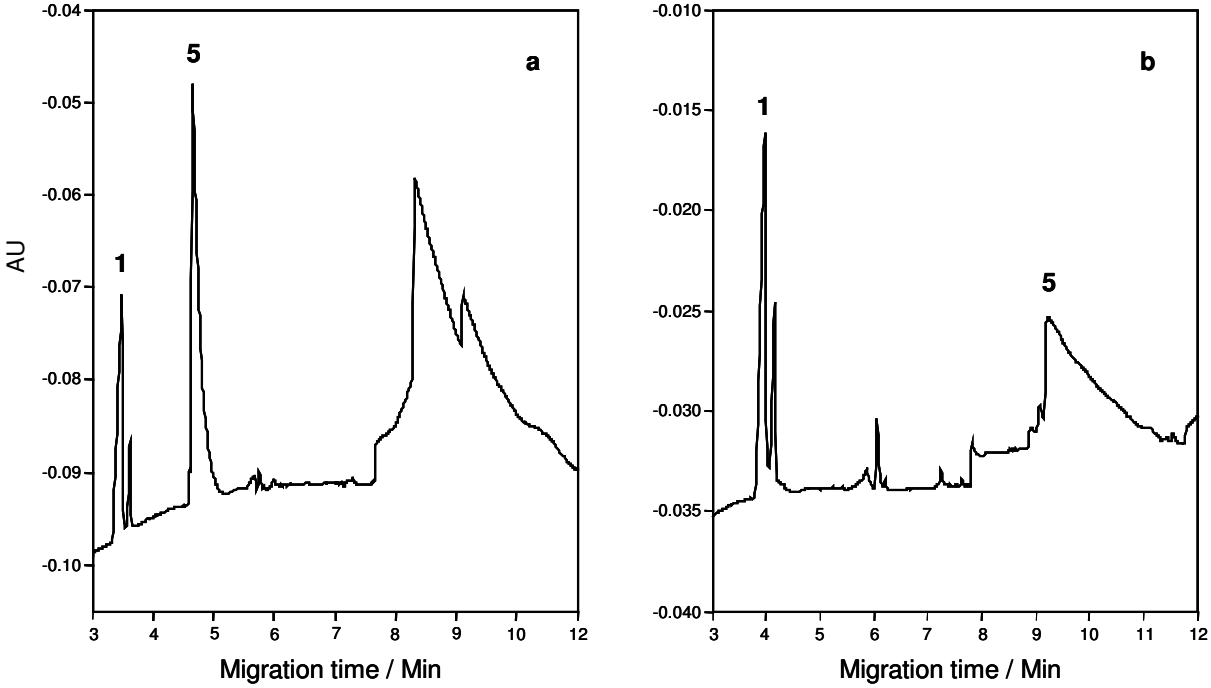


Fig. 4.5 The venom fraction CSG-1, spiked with Cl^-, separated in the commercial Fluka buffer (**a**), and with additional 0.1 mM of Ca^{2+} in the buffer (**b**), in negative ion mode. For conditions and components see Fig. 4.4.

Positive ions

The system was tested with a mixture of five inorganic cations: ammonium, sodium, potassium, calcium, and magnesium, which were properly separated.(Fig. 4.6a) Histamine gave a negative peak under the conditions listed above.(Fig. 4.6b)

All the venom fractions contained ammonium and sodium. The histamine content of CSG-4 and CSG-6 could be confirmed with the method.(Fig. 4.7a,b)

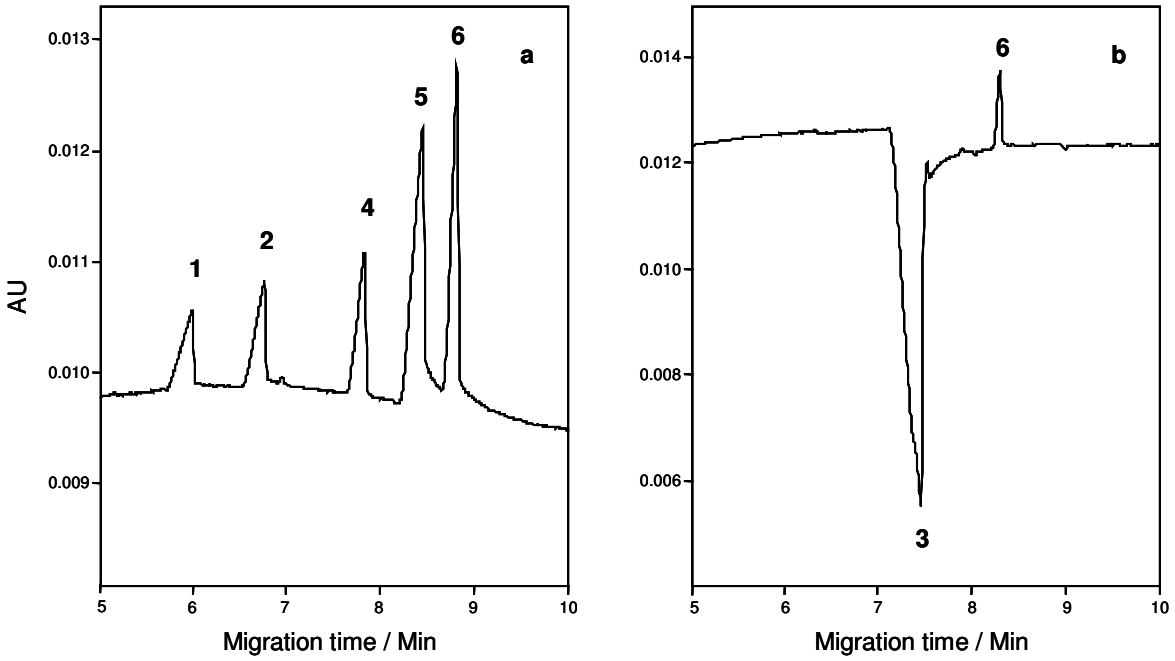


Fig. 4.6 Separation of the inorganic cation standard (**a**) and histamine spiked with Mg^{2+} (**b**) in positive ion mode. Components: (**1**) NH_4^+, (**2**) K^+, (**3**) histamine, (**4**) Na^+, (**5**) Ca^{2+}, and (**6**) Mg^{2+}. Injection: 0.5 psi, 5s; separation: Fluka cation buffer, 10 kV; detection: UV indirect at 217 nm.

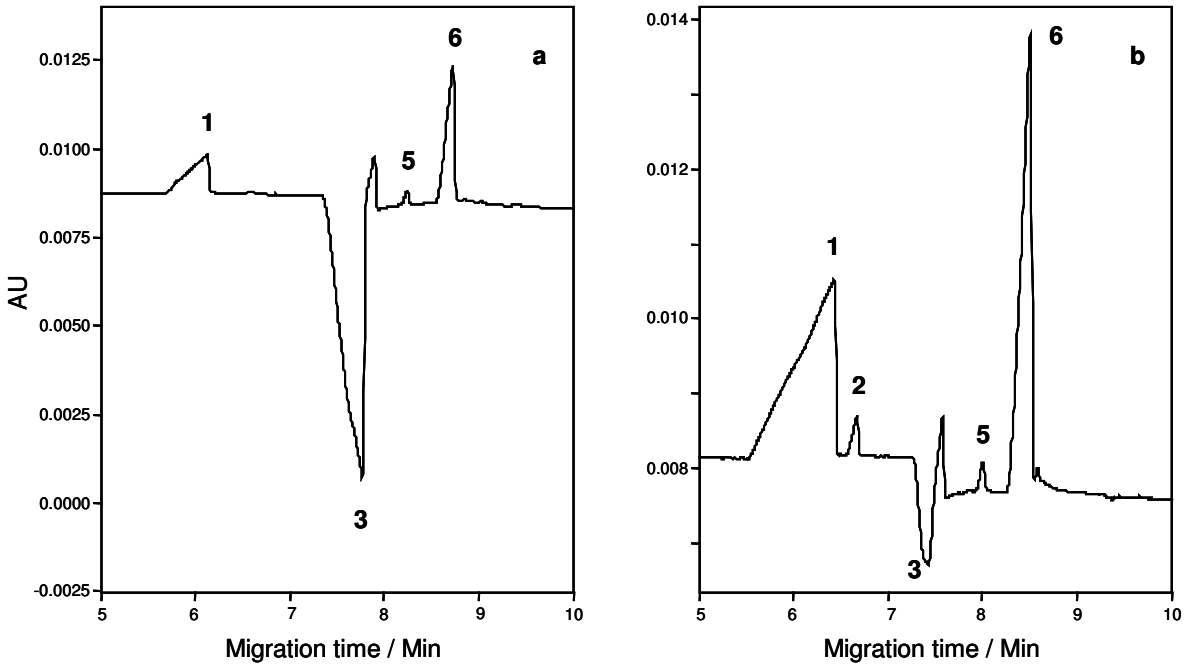


Fig. 4.7 The venom fractions CSG-4 (**a**) and CSG-6 (**b**) spiked with Mg^{2+}, separated in positive ion mode. For conditions and components see Fig. 4.6.

4.4 Discussion

The results of the CE experiments confirmed the findings of the GC-MS analysis; citric acid and histamine are LMW constituents of the venom of *C. salei*. Citric acid was found in fractions CSG-1 and CSG-3, which were the first early eluting peaks in the RP-HPLC chromatogram of the gel chromatography fractions D001/3c and D001/4a, respectively. (Scheme 4.1) The related fraction from D001/4b, CSG-5, was probably more diluted, since no peak was observed by the GC-MS analysis. Histamine was the main component in samples CSG-4 and CSG-6, the second early eluting fractions from D001/4a and D001/4b, respectively. Again, the same fraction of D001/3c, CSG-2, was probably not concentrated enough.

No aconitic acids were detected in fractions CSG-1 and CSG-3 by CE-UV/DAD, although the corresponding GC peak was significant. Aconitic acid is produced probably in the GC injector by dehydration of citric acid, since its occurrence seems to correlate with that of citric acid in the samples (Table 4.2).

Several substances were detected by CE-UV/DAD and declared unknown due to lack of identification and reference compounds. Nevertheless, all the fractions contained late eluting peaks, probably phosphate and acetate, the latter from sample preparation. Similarly, ammonium was detected in each fraction. The samples CSG-1, CSG-3, and CSG-5 contained most likely chloride as early eluting component, whereas sodium, potassium and calcium were also found in different concentrations in the fractions.

Although volatility of analytes can be enhanced by derivatization of polar functional groups, gas chromatographic methods are not suitable for the analysis of thermally unstable compounds, for example the acylpolyamines from spider venom. In addition, derivatization may lead to by-products, the possibility to gain impurities during sample preparation is higher.

It is more convenient to analyze a biological sample, without transferring it to an organic solvent, by HPCE methods. Nevertheless, identification of the components is only possible by offline or online MS detection. The direct coupling of CE with MS was first reported in 1987[120] and has been intensively investigated in the last years.[121-124] Still, one condition has to be fulfilled: replacing non-volatile BGE constituents for volatile ones, which may be challenging to do without loss of separation efficiency. Ammonium formate-formic acid and ammonium acetate-acetic acid buffers are used in the acidic and neutral pH range, whereas

separation of basic compounds can be achieved with ammonium bicarbonate (basic conditions).[125,126] In sheath liquid interfaces the capillary effluent is diluted with methanol or acetonitril.

4.5 Summary

Two known constituents of spider venom: citric acid and histamine have been identified by GC-MS and CE-UV/DAD in the LMW fractions of the venom of *Cupiennius salei*. Citric acid inhibits self-intoxication by building a complex with calcium, which is necessary for the activity of the phospholipases.[91] Histamine is a spreading factor; it facilitates toxin transport into the tissue by damaging capillary walls.[99]

Inorganic ions, as well as several unidentified compounds were also detected by CE-UV/DAD. In order to complete the investigation of the venom of *Cupiennius salei* CE-MS analysis of the LMW fractions would be necessary.

5 EXPERIMENTAL

5.1 Materials

Chromatographic analysis

Dichloromethane Suprasolv (MERCK), made for gas chromatographic trace analysis, was used for analytical experiments in non-aqueous media was.

Derivatization of natural samples prior GC analysis was carried out with *N*-methyl-*N*-trimethylsilyltrifluoracetamide (MSTFA) (CE Chromatography Service). For methylation the methanolic solution of trimethyl sulfonium hydroxide 0.25M (TMSH) (Fluka) or a self-made solution of diazomethane in dichloromethane was used (see section 5.3).

HPLC grade hexane (ACROS or ROTH) was applied as mobile phase for HPLC separations during isolation of the C35 terpenes.

Capillary electrophoresis experiments

Deionized water of CE grade (Fluka) was used to prepare standard solutions, as well as to dilute venom samples. Stock solutions of both organic and inorganic ions and solutions for rinsing the capillary were made from substances of analytical purity. The concentration of the stock solutions was 100 mM. Commercially available running buffers from Fluka were applied for the separation in anionic and cationic mode (section 4.3.2.2).

Synthesis

Reagents were purchased from different firms: ruthenium trichloride from Aldrich, Amberlyst from ACROS, sodium periodate from JANNSEN CHIMICA, periodic acid from MERCK, and (-)/(+)-citronellene from Fluka. Solvents were only distilled before use. Distilled water was used as reaction solvent or solvent in work-up procedures. Reactions were followed with thin-layer chromatography (TLC) on Polygram SIL G/UV_{254} plates from Machrey&Nagel; substances were detected with a UV lamp at 254 nm by immersion into the charring reagent. Flash gel chromatography was made on silica M60 (0.04-0.063 mm, 230-400 mesh ASTM) from Machrey&Nagel under pressure.

5.2 Instrumentation and equipment

Derivatization

Sample concentration and derivatization was carried out in a heating block (Lab-Line Instruments Inc.). A gentle stream of nitrogen was applied to evaporate residual solvents and by-products.

Gas chromatographic analysis

Gas chromatographic measurements were performed on a GCTop 8000 gas chromatograph, connected to an AS800 autosampler from ThermoQuest, with hydrogen as carrier gas. Flame ionization detection was used with nitrogen as make-up gas. The carrier flow was usually set to 1 ml/min at 100°C, or else see actual description in the text. Data acquisition and evaluation was carried out by the software ChromCard 1.19 from ThermoQuest.

Derivatized marine extracts were analyzed on the same BPX-5 (25 m x 0.25 mm x 0.25 µm) capillary from SGE Inc. under identical conditions. The split/splitless inlet was in splitless mode at a temperature of 250°C (split ratio 1:20 with valve time set to 0.75 min) and the FID operated at 280°C. The initial oven temperature was 50°C for 2 minutes followed by a slow ramp of 3 °C/min to the final temperature of 320°C. Retention time shifts were followed by regular analysis of a hydrocarbon standard mixture containing C11-C44 alkanes.

Routine measurements as well as the analysis of reaction mixtures were carried out on an other BPX-5 (25 m x 0.25 mm x 0.25 µm) capillary with the inlet in split mode (1:20) and a normal ramp of 10 °C/min. Other parameters were identical to that described before.

GC-MS analysis

A GC 6890 was coupled to an MSD 5973 from Hewlett Packard to identify substances separated on a BPX-5 (25 m x 0.25 mm x 0.25 µm) capillary from SGE Inc. with helium as carrier gas and a constant flow rate of 1 ml/min. The split/splitless inlet was operated either in split (1:20) or in splitless mode (1:20, 0.75 min valve time) at 250°C. Samples were injected by a HP G1513A autosampler or manually. The oven program was the same as for GC.

Heated zones in the MS were set at 320°C, 230°C and 150°C for the transfer line, ion source and the quadrupoles, respectively. The ion source was used in EI+ mode at 70 eV. Zone temperature and final oven temperature values were lower when using a chiral capillary Data was recorded by the HP software ChemStation v. A.03.00, whereas data evaluation was carried out with various programs, such as MassLib, ChemStation Enhanced Data Analysis (HP) and NIST98.

GC-IR analysis

Infrared spectra of selected compounds were obtained on a HP GC 6890 system coupled with a HP 5965A infrared detector. Helium was the carrier gas at a constant flow of 2 ml/min. The capillary HP5 (30 m x 0.32 mm x 0.25 μm) was used. Samples were injected splitless (1:20, valve time 1 min). Data acquisition was performed by the software GRAMS/32 v. 4.14.

HPLC analysis

High-performance liquid chromatographic experiments were made on a Spectra System from ThermoQuest containing a P4000 pump, an AS3000 autosampler and an SN4000 diode array detector. A Superspher Si60 column (250 x 4 mm, 4 μm) from MERCK was used to separate the compounds with hexane as the mobile phase. Flow rate was set to 1 ml/min. Fractions were collected manually. ChromQuest 2.1 was used for data acquisition and analysis.

Capillary electrophoresis experiments

Spider venom samples and standards were analyzed on a P/ACE MDQ system from Beckman Coulter. The UV detector had a deuterium lamp and four filters with a wavelength of 200, 214, 254 and 280 nm. The photo-diode-array (PDA) detector contained a deuterium and a mercury lamp providing a wavelength range of 190-600 nm. Measurement data were recorded and analyzed by the software 32Karat 4.0 (Beckman). An uncoated fused silica capillary 60.2 cm x 0.75 mm i.d. (Beckman) was used with self-burned optical window at 10 cm from the outlet end. The capillary was maintained at 20 or 25°C, the samples at 20°C during measurements. The coolant liquid was a mixture of perfluoro compounds (C5-C18) from Beckman.

Nuclear magnetic resonance spectroscopy

NMR spectra of synthesis products were obtained on a AC-200F (^{1}H 200 MHz, ^{13}C 50 MHz), for natural products on a DRX-400 (^{1}H 400 MHz, ^{13}C 100 MHz) instrument from BRUKER. The internal standard was tetramethylsilane.

5.3 Procedures

5.3.1 Derivatization methods

Trimethylsilylation of marine extracts

Samples were dissolved in dichloromethane to get similarly concentrated solutions, and transferred into a 1.5 ml-vial (CE Chromatography Service). 50 µl of the solution was mixed with 100 µl of MSTFA in a micro-insertion vial, closed tightly and maintained at 50°C for an hour. The rest of MSTFA and the solvent was evaporated at 50°C under a gentle stream of nitrogen. Samples were redissolved in dichloromethane.

Methylation A

TMSH converts aromatic alcohols to the methyl ether and acids as well as esters to the methyl ester. To 50 µl of a sample 100 µl of the TMSH solution was added. The vial was closed tightly and kept at room temperature for 15 minutes. The rest of TMSH and the solvent was evaporated at 50°C under a gentle stream of nitrogen. The sample was redissolved in dichloromethane.

Methylation B

A self-made diazomethane solution was applied to derivatize organic acids in natural samples. Solution A was prepared from 1.71 g (8 mmol) Diazald (*N*-methyl-*N*-nitroso-p-toluolsulfonamide) dissolved in 5 ml of a 1:1 mixture of diethyleneglycol-dimethylether :diethyl ether. 10 g of KOH was dissolved in a 1:1 mixture of water and methanol to give solution B. A volume of 0.5 ml of both solutions was added to a GC vial. The gaseous

diazomethane was led through a buffer vial into 1 ml of dichloromethane in a third vial, which was cooled with ice. The vials were connected with teflon tubing, and a canulla served as an opening in the collection vial. Yellow coloring of the diazomethane solution showed saturation. It was stored in a refrigerator at 4°C for a week.

Samples were titrated with small portions of the diazomethane solution until it became colorless. In the case of colorful samples an excess of diazomethane was used. The rest of diazomethane and the solvent was evaporated at 50°C under a gentle stream of nitrogen. Samples were redissolved in dichloromethane.

Determination of double bond position

The sample was dissolved in dichloromethane. To 50 µl of the solution was added 50 µl of dimethyl disulfide and 5 µl of 5% etherous iodine solution. The closed vial was kept at 60°C for a day. The cold mixture was diluted with 200 µl of dichloromethane. Iodine in excess was reduced by shaking the mixture with 200 µl of aqueous sodium sulfite solution. After separation of the phases, the solvent in the organic phase was evaporated in a gentle stream of nitrogen. The residue was redissolved in dichloromethane.

Catalytic hydrogenation of double bonds in micro-scale

The lipophilic sample was dissolved in 0.5 ml of isooctane in a GC vial with screw cap and a small magnetic stirrer. The experiment was carried out at room temperature under continuous stirring. A tiny portion of Pd/C (10%) was added with a spatula edge and the mixture was stirred for 10 minutes. Hydrogen was introduced into the vial via stainless steel tubing with a valve and a canulla at the end. A second canulla provided gas purge through the septum. The pressure was maintained at 0.1 bar for 45 minutes before removing the extra canulla. The hydrogen valve was closed and the reaction was stirred for another 15 minutes under hydrogen atmosphere. The mixture was filtrated through silica in a Pasteur-pipette, and the filtrate was concentrated under a gentle stream of nitrogen at 50°C. The residue was dissolved in dichloromethane.

5.3.2 Oxidative cleavage

(*R*)-2-methyl-glutaric acid dimethyl ester (29a)[50,59]

A solution of 1.73 g (8.1 mmol) of sodium-periodate ($NaIO_4$) in 3 ml of water was stirred at room temperature for 20 minutes, which was followed by the addition of 2 ml of acetonitrile, 2 ml of carbon tetrachloride and 180 µl (140 mg, 1 mmol) of (-)-(R)-citronellene. By introducing 4 mg (0.02 mmol, 2 mol%) of ruthenium trichloride hydrate ($RuCl_3.(H_2O)_n$), the temperature was kept between 20-80°C with a water bath and the mixture became soon orange-brown. The reaction was followed with TLC and GC after micro-derivatization with TMSH. After 4 hours of vigorous stirring 10 ml of diethyl ether was added which caused a rush darkening of the mixture. The aqueous phase was acidified with 2N HCl solution and extracted 3 times with diethyl ether. The combined organic phases were washed with brine, dried over magnesium sulfate ($MgSO_4$) and concentrated at 50°C under normal pressure. The residue was filtrated through a celite pad in vacuum (water pump), and the solvent was removed in vacuum. The crude product was used without purification for the next reaction.

Methylation: The crude product was dissolved in 8 ml of methanol and 0.32 g of Amberlyst 15 was added. The mixture was stirred vigorously at room temperature for a day. The resin was filtrated off and washed with diethyl ether. The mixture was concentrated in vacuum. The crude product was purified with flash chromatography on silica to give 10 mg (yield 6%) product.

Spectroscopic data (MS, NMR) was identical to that in the literature.[127,128]

(*S*)-2-methyl-glutaric acid dimethyl ester (29b)

The same procedure was used as for the (*R*)-enantiomer, but with 180 µl (1 mmol) of (+)-(*S*)-citronellene. The crude product was purified with flash chromatography on silica to give 9 mg (yield 5%) product.

Spectroscopic data (MS, NMR) was identical to that in the literature.[127,128]

Oxidative degradation of the natural product 26

A mixture containing 0.5 mg (1 µmol) of the isolated natural product **26**, 8 mg (35 µmol, 35 equiv.) of H_5IO_6, 20 µl of acetonitrile, 20 µl of carbon tetrachloride, and 30 µl of distilled water was stirred 5 minutes in a GC vial at room temperature. Then a

minute amount of $RuCl_3.(H_2O)_n$ was introduced and the whole mixture was stirred 5 hour long. After shaking with 200 µl of dichloromethane (Suprasolv, MERCK) and 200 µl of brine, the phases were separated. The organic phase was concentrated in a gentle stream of nitrogen at 30°C and derivatized with a solution of diazomethane. The solvent was evaporated and the residue was dissolved in dichloromethane.

6 ABBREVIATIONS

2D	two dimensional
ACP	acyl carrier protein
APCI	atmospheric pressure chemical ionization
BGE	background electrolite
BSTFA	*N*,*O*-bis(trimethylsilyl)trifluoracetamide
CA	cluster analysis
CE	capillary electrophoresis
d	doublet
DB	double bond
DBE	double bond equivalent
DMDS	dimethyldisulfide
EI	electron impact
EOF	electroosmotic flow
ESI	electrospray
Et	ethyl
FA	fatty acid
FAS	fatty acid synthase
FID	flame ionization detector
FPP	farnesyl pyrophosphate
GC	gas chromatography
GC-HRMS	gas chromatography high-resolution mass spectrometry
GC-MS	gas chromatography mass spectrometry
GGPP	geranylgeranyl pyrophosphate
HBI	highly-branched isoprenoid
HCA	hierarchical cluster analysis
HMBC	hetero multiple bond correlation
HPCE	high-performance capillary electrophoresis
HPLC	high-performance liquid chromatography
HPLC-MS	high-performance liquid chromatography mass spectrometry

HR-MS	high-resolution mass spectrometry
HSQC	heteronuclear single quantum coherence
i.d.	internal diameter
J	coupling constant
LMW	low-molecular weight
m	multiplet
M	molecular ion
Me	methyl
MSTFA	*N*-methyl-*N*-trimethylsilyltrifluoroacetamide
m/z	mass per charge ratio
NMR	nuclear magnetic resonance
PCA	principle component analysis
PDA	photo diode-array
Pd/C	palladium on activated charcoal
PKS	polyketide synthase
R	alkyl chain
RI	retention index
s	singulet
SIM	single ion monitoring
SPSS	statistical software
t	tirplet
TIC	total ion chromatogram
TMS	trimethylsilyl
TMSH	trimethylsulfonium hydroxide
UV	ultraviolet

7 REFERENCES

1 D. J. Newman, G. M. Cragg, K. M. Snader, *Nat. Prod. Rep.* **2000**, *17*, 215-234.

2 J. Mann, Secondary Metabolism, (Ed.: P. W. Atkins, J.S.E. Holker, A. K. Holliday), Clarendon Press, Oxford, 1987, p. 1-23.

3 W. Francke, Signalstoffe: Von kleinen und großen Molekülen, Lecture Abstract, BioPerspectives, May 4-6, 2004, Wiesbaden, Germany.

4 N. Olsen, T. M. Swanson, R. A. Luxmoore, WCMC Biodiversity Series No. 7, (Ed.: T. M. Swanson, R. A. Luxmoore), World Conservation Press, Cambridge, 1997, p. 57-76.

5 WCMC Biodiversity Series No. 4, (Ed.: B. Groombridge, M. D. Jenins), World Conservation Press, Cambridge, 1996, p. 3-27.

6 J. W. Blunt, Brent R. Copp, M. H. G. Munro, P. T. Northcote, M. R. Prinsep, *Nat. Prod. Rep.* **2004**, *21*, 1-49.

7 A. M. S. Mayer, M. T. Hamann, *Marine Biotechnology* **2004**, *6*, 37-52.

8 R. C. Willis, *Modern Drug Discovery* **2002**, *5*, 32-38.

9 F. Pietra, A secret world: natural products of marine life, Birkhäuser Verlag, Boston, 1990, p. 1-9, 65-78.

10 E. Winterfeldt, Marine Natural products - Feeding Data Bases or Triggering Innovations, Hellmich PrePress & Print GmbH, 1999, Berlin.

11 a) T. Higa, Marine natural Products, Vol. IV, (Ed.: P. J. Scheuer), Academic press Inc., New York, 1981, p. 93-145, b) M. Barbier, p. 147-186.

12 W. Fenical, Assessing Marine Microbial Biodiversity for Drug Discovery, Lecture Abstract, 3rd European Conference on Marine Natural Products, September 15-20, 2002, Elmau, Germany.

13 R. H. Feling, G. O. Buchanan, T. J. Mincer, C. A. Kauffman, P. R. Jensen, W. Fenical, *Angew. Chem.* **2003**, *115*, 369-371.

14 R. I. Amann, W. Ludwig, K. H. Schleifer, *Microbiol. Rev.* **1995**, *59*, 143-169.

15 E. K. Wilson, Science and Technology, January 20, 2003, p. 37-38.

16 S. Chesnov, L. Bigler, M. Hesse, *Helv. Chim. Acta* **2001**, *84*, 2178-2197.

17 S. Chesnov, L. Bigler, M. Hesse, *Eur. J. Mass Spectrom.* **2002**, *8*, 1-16.

18 Dissertation of Dr. Markus Müller.

19 Dissertation of Dr. Katja Stritzke.

20 K. Stritzke, S. Schulz, H. Laatsch, E. Helmke, W. Beil, *J. Nat. Prod.* **2004**, *67*, 395-401.

21 D. J. Robins, M. A. Sefton, *Phytochemistry* **1984**, *23*, 200-201.

22 Y. S. Bahn, J.-M. Park, D. H. Bai, S. Takase, J. H. Yu, *J. Antibiot.* **1998**, *51*, 902-907.

23 K. Varmuza, *Anal. Chem. Symposia Series* **1983**, *15*, 19-34.

24 a) The elements of statistical learning ,(Ed.: T. Hastie, R. Tibshirani, J. Friedman), Springer -Verlag New York Inc, 2001, p. 1-2, b) p. 437-503.

25 a) Pattern classification, (Ed.: R. O. Duda, P. E. Hart, D. G. Stork), John Wiley & Sons, Inc., New York, 2001, p. 187-188, b) p. 552-559.

26 Cluster Analysis, www.statsoft.com/textbook/stcluan.html#a.

27 B. K. Lavine, Clustering and Classification of Analytical Data, Encyclopedia of Analytical Chemistry, (Ed.: R. A. Meyers), Vol.11, John Wiley and Sons Ltd, Chicester, 2000, p. 1-21 (www.wiley.co.uk/wileychi/eac/pdf/A5204-W.PDF).

28 Intelligent Data Analysis, (Ed.: M. Berthold, David J. Hand), Springer, 2003, p. 99-127.

29 B. R. Kowalski, *Anal. Chem.* **1975**, *47*, 1152A-1162A.

30 Chemometrics in Chromatography, www.infometrix.com/apps/14-0296_ ChromatogAO.pdf.

31 I. Miksik, A. Eckhardt, T. Cherháti, E. Forgács, J. Zicha, Z. Deyl, *J. Chrom. A* **2001**, *921*, 81-91.

32 R. J. Mashall, R. Turner, H. Yu, E. H. *Cook, J. Chrom.* **1987**, *297*, 235-244.

33 P. J. Dunlop, C. M. Bignell, D. B. Hibbert, *Aust. J. Bot.* **1997**, *45*, 1-13.

34 K. J. Johnson, R. E. Synovec, *Chemometrics and Intelligent Laboratory Systems* **2002**, *60*, 225-237.

35 www.spss.com/corpinfo/history.htm

36 J. Janssen, W. Laatz, Statistische Datenanalyse mit SPSS für Windows, Sringer, 2003.

37 A. L. Duran, J. Yang, L. Wang, L. W. Sumner, *Bioinformatics* **2003**, *19*, 2283-2293.

38 K. J. Johnson, B. W. Wright, K.H. Jarman, R. E. Synovec, *J. Chrom. A* **2002**, *996*, 141-155.

39 M. L. McConnell, G. Rhodes, U. Watson, M. Novotný, *J. Chrom.* **1979**, *162*, 495-506.

40 R. P. Evershed, Handbook of Derivatives for Chromatography, (Ed.: K. Blau, J. M. Halket), John Wiley & Sons Ltd, New York, 1993, p. 52-108.

41 K. Yamauchi, T.Tanabe, M. Kinoshita, *J. Org. Chem.* **1979**, *44*, 638-639.

42 W. Butte, *J. Chrom.* **1983**, *261*, 142-145.

43 D. Joulain, W. A. König, The atlas of spectral data of sesquiterpene hydrocarbons, E.B.-Verlag, Hamburg, 1998, p. 253-256

44 A. I. Meyers, D. Stoinova, *J. Org. Chem.* **1997**, *62*, 5219.

45 T. A. van Beek, G. P. Lelyveld, *Phytochem. Anal.* **1991**, *2*, 26.

46 T. D. H. Bugg, Ch. J. Winfield, *Nat. Prod. Rep.* **1998**, 513-524.

47 N. C. Deno, B. A. Greigger, L. A. Messer, M. D. Meyer, S. G. Stroud, *Tetrahedron Lett.* **1977**, *20*, 1703-1704.

48 O. Neunhoeffer, *Journal f. prakt. Chemie* **1932**, *133*, 99,106.

49 H. Klein, A. Steinmetz, *Tetrahedron Lett.* **1975**, *48*, 4249-4250.

50 H. J. Carlsen, T. Katsuki, V. S. Martin, K. B. Sharpless, *J. Org. Chem.* **1981**, *46*, 3939-3940.

51 R. Liotta, W. S. Hoff, *J. Org. Chem.* **1980**, *45*, 2887-2890.

52 M. Ono, K. Suzuki, S. Tanikawa, H. Akita, *Tetrahedron: Asymmetry* **2001**, *12*, 2597-2604.

53 Y. Liu, J. Hong, Ch-O. Lee, K. S. Im, N. D. Kim, J. S. Choi, J. H. Jung, *J. Nat. Prod.* **2002**, *65*, 1307-1314.

54 R. Brückner, Reaktionmechanismen: organische Reaktionen, Stereochemie, moderne Synthesemethoden, Spektrum Akademische Verlag, Heidelberg, Berlin, Oxford, 1996, p. 505-509.

55 J. Frunzke, Ch. Loschen, G. Frenking, *J. Am. Chem. Soc.* **2004**, *126*, 3642-3652.

56 C. Djerassi, R. R. Engle, *J. Am. Chem. Soc.* **1953**, *75*, 3838-3840.

57 U. A. Spitzer, D. G. Lee, *J. Org. Chem.* **1974**, *39*, 2468-2469.

58 M. T. Nunez, V. S. Martin, *J. Org. Chem.* **1990**, *55*, 1928-1932.

59 M. Petrini, R. Ballini, E. Marcantoni, *Syn. Comm.* **1988**, *18*, 847-853.

60 S. Schulz, *Lipids* **2001**, *36*, 637-647.

61 J. L. Mieloszynski, O. Aberkane, D. Paquer, *Sulfur Letters* **1989**, *9*,17-21.

62 G. Holzmann, R. Susilo, R. Gmelin, *Org. Mass Spec.* **1982**, *17*, 165-172.

63 J.-X. Zhang, J. Ni, X.-J. Ren, L. Sun, Z.-B. Zhang, Z.-W. Wang, *Chem. Senses* **2003**, *28*, 381-388.

64 J.-X. Zhang, L. Sun, Z.-B. Zhang, Z.-W. Wang, Y. Chen, R. Wang, *J. Chem. Ecol.* **2002**, *28*, 1287-1297.

65 P. Dubs, M. Joho, *Helv. Chim. Acta* **1978**, *61*, 1404-1406.

66 M. Ritzau, M. Keller, P. Wessels, K. O. Stetter, A. Zeeck, *Liebigs Ann. Chem.* **1993**, 871-876.

67 H.-P. Kruse, H. Anger, *Journal f. prakt. Chemie* **1981**, *323*, 169-173.

68 T. Kawai, Y. Ishida, *J. Agric. Food Chem.***1989**, *37*, 1026-1031.

69 Y. Zhang, C.-T. Ho, *J. Agric. Food Chem.***1989**, *37*, 1016-1020.

70 S.-S. Hwang, J. T. Carlin, Y. Bao, G. J. Hartman, C.-T. Ho, *J. Agric. Food Chem.* **1989**, *34*, 538-542.

71 G. Vernin, C. Boniface, J. Metzger, T. Obretenov, J. Kantasubrata, A. M. Siouffi, J. L. Larice, D. Fraisse, *Bull. Soc. Chim. France* **1987**, *4*, 681-694.

72 J. Dickschat, T. Martens, T. Brinkhoff, S. Schulz, *Appl. Environ. Microbiol.*, in preparation.

73 S. D. Bull, S. G. Davies, R. M. Parkin, F. Sámchez-Sancho, *J. Chem. Soc., Perkin Trans. 1* **1998**, 2313-2320.

74 H. Ott, A. J. Frey, A. Hofmann, *Tetrahedron* **1963**, *9*, 1675-1684.

75 C. Eguchi, A. Kakuta, *J. Am. Chem. Soc.* **1974**, *96*, 3985-3989.

76 M. J. O. Anteunis, *Bull. Soc. Chim. Belg.* **1978**, *87*, 627-650.

77 P. Gund, D. F. Veber, *J. Am. Chem. Soc* **1979**, *101*, 1885-1887.

78 E. Breitmaier, Terpene: Aromen, Düfte, Pharmaka, Pheromone, B. G. Teubner Stuttgart Leipzig, 1999, p. 25-28.

79 K. H. C. Baser, Pure Apl. Chem. **2002**, *74*, 527-545.

80 B. S. Rao, J. L. Simonsen, *J. Am. Chem. Soc.* **1928**, 2496-2505.

81 G. F. Russel, W. G. Jennings, *J. Agric. Food Chem.* **1969**, *17*, 1107-1112.

82 A. F. Halim, R. P. Collins, *J. Agric. Food Chem.* **1975**, *23*, 506-510.

83 B. P. Manquin, J. A. Morgan, J. Ju, T, Müller-Späth, D. S. Clark, *Extremophiles* **2004**, *8*, 13-21.

84 S. J. Hird, R. Evens, S. J. Rowland, *Marine Chemistry* **1992**, *37*, 117-129.

85 S. T. Belt, D. A. Cooke, S. J. Hird, S. Rowland, *J. Chem. Soc., Chem. Commun.* **1994**, 2077-2078.

86 S. T. Belt, D. A. Cooke, J.-M. Robert, S. Rowland, *Tetrahedron Lett.* **1996**, *37*, 4755-4758.

87 S. T. Belt, D. W. G. Allard, G. Massé, J.-M. Robert, S. Rowland, *Tetrahedron Lett.* **2001**, *42*, 5583-5585.

88 J.-F. Rontani, A. Mouzdahir, V. Michotey, P. Caumette, P Bonin, *Appl. Environ. Microbiol.* **2003**, 4167-4176.

89 T. Koyama, K. Ogura, Comprehensive Natural Products Chemistry, Vol. 2, (Ed.: D. E. Cane), Elsevier, New York, 1999, p. 69-96.

89b M. L. Wise, R. Croteau, Comprehensive Natural Products Chemistry, Vol. 2, (Ed.: D. E. Cane), Elsevier, New York, 1999, p. 97-154.

90 D. O'Hagan, The Polyketide Metabolites, Ellis Horwood Ltd., London, 1991, p. 19-43.

91 S. Schulz, *Angew. Chem. Int. Ed. Engl.*, **1997** *36*, 314-326.

92 M. Ori, H. Ikeda, *J. Toxicology. Toxin reviews* **1998**, *17*, 405-426.

93 Y. Itagaki, T. Nakjima, *J. Toxicol.-Toxin Reviews* **2000**, *19*, 23-52.

94 L. Kuhn-Nentwig, J. Schaller, W. Nentwig, *Toxicon* **2004**, *43*, 543-553.

95 G. J. Binford, *Toxicon* **2001**, *39*, 955-968.

96 A. Horni, D. Weickmann, M. Hesse, *Toxicon* **2001**, *39*, 425-428.

97 K. D. McCormick, J. Meinwald, *J. Chem. Ecol*. **1993**, 2411-2451.

98 L. Kuhn-Nentwig, J. Schaller, W. Nentwig, *Toxicon* **1994**, *32*, 287-302.

99 L. Kuhn-Nentwig, A. Bücheler, A. Studer, W. Nentwig, *Naturwissenschaften* **1998**, *85*, 136-138.

100 J. Meinwald, Th. Eisner, *Proc. Natl. Acad. Sci. USA* **1995**, *99*, 14-18.

101 J. McCormick, Y. Li, K. McCormick, H. I. Duynstee, A. K. van Engen, G. A. van der Marel, B. Ganem, J. H. van Boom, J. Meinwald, *J. Am. Chem.Soc*. **1999**, *121*, 5661-5665.

102 C.-H. Lo, S.-H. Chiou, *J. Chrom. B* **1990**, *530*, 129-136.

103 N. Mokrzycki, H. Drobecq, H. Gras, A. Tartar, *J. Chrom. A* **1999**, *839*, 57-69.

104 J. R. Perkins, K. B. Tomer, *Anal. Chem.* **1994**, *66*, 2838-2840.

105 J. R. Perkins, K. B. Tomer, Eur. J. Biochem. **1995**, *233*, 815-827.

106 a) E. Kenndler, High-Performance Capillary Electrophoresis, (Ed.: M. G. Khaledi), John Wiley ans Sons, Inc., New York, 1998, p. 58-69, b) B. L. Karger, p. 3-24.

107 R. Weinberger, Practical Capillary Electrophoresis, Academic Press, San Diego, 2000, a) p. 25-72, b) p. 1-24, c) p. 368, d) p. 365-422, e) p. 423-424, f) p. 99-125, g) p. 166-196, h) p. 265-284, i) p. 234-241, j) p. 335-360, k) p. 309-316.

108 A. Hartl, www.wissenschaft-online.de/artikel/619918&template=bild_popup&_ bild=1.

109 B. Tenberken, K. Bächmann, *J. Chrom. A* **1997**, *775*, 372-377.

110 M. Arellano, P. Jomard, S. E. Kaddouri, C. Roques, F. Nepveu, F. Couderc, *J. Chrom. B* **2000**, *741*, 89-100.

111 Z. Chen, C. Tang, J. C. Yu, *J. High Resol. Chromatogr.* **1999**, *22*, 379-385.

112 J. Dahlén, J. Hagberg, S. Karlsson, *Fresenius J. Anal. Chem.* **2000**, *366*, 488-493.

113 M. M. Rhemrev-Boom, *J. Chrom. A* **1994**, *680*, 675-684.

114 H. Sirén, R. Kokkonen, T. Hiissa, T. Särme, O. Rimpinen, R. Laitinen, *J. Chrom. A* **2000**, *895*, 189-196.

115 L. Arce, A. Rios, M. Valrárcel, *Chromatographia* **1997**, *46*, 170-176.

116 W. Beck, H. Engelhardt, *Chromatographia* **1992**, *33*, 313-316.

117 R. Zhang, C. L. Cooper, Y. Ma, *Anal. Chem.* **1993**, *65*, 704-706.

118 G. Zhou, Q. Yu, Y. Ma, J. Xue, Y. Zhang, B. Lin, *J. Chrom. A* **1995**, *717*, 345-349.

119 X. Xiong, S. F. Y. Li, *J. Chrom. A* **1998**, *822*, 125-136.

120 J. A. Olivares, N. T, Nguyen, C. R. Yonker, R. D. Smith, *Anal. Chem.* **1987**, *59*, 1232-1236.

121 P. Schmitt-Kopplin, M. Frommberger, *Electrophoresis* **2003**, *24*, 3837-3867.

122 C. W. Huck, G. Stecher, R. Bakry, G. K. Bonn, *Electrophoresis* **2003**, *24*, 3977-3997.

123 S. A. Shamsi, *Electrophoresis* **2002**, *23*, 4036-4051.

124 A. von Brocke, G. Nicholson, E. Bayer, *Electrophoresis* **2001**, *22*, 1251-1266.

125 T. E. Wheat, K. A. Lilley, J. F. Banks, *J. Chrom. A* **1997**, *781*, 99-105.

126 J. McCourt, G. Bordin, A. R. Rodrígez, *J. Chrom. A* **2003**, *990*, 259-269.

127 A. L. Wilkins, Y. Lu, S.-T. Tan, *J. Agric. Food Chem.* **1995**, *43*, 3021-3025.

128 S. Tyrlik, *J. Organomet. Chem.* **1973**, *59*, 365-378.

8 APPENDIX

8.1 Mass spectra

8.1.1 Artifacts

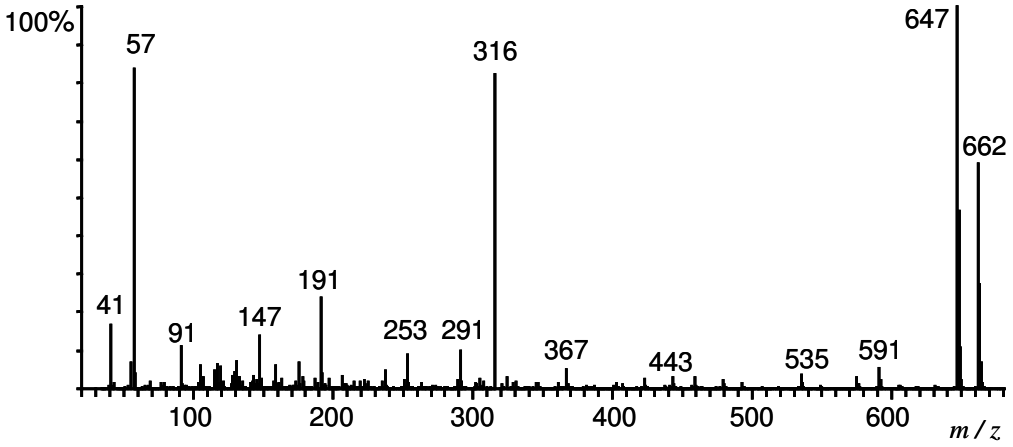


Fig. 8.1 Tris-(2,4-di-tertbutylphenyl)-phosphate (**19a**), M662, from 1/Hel39.

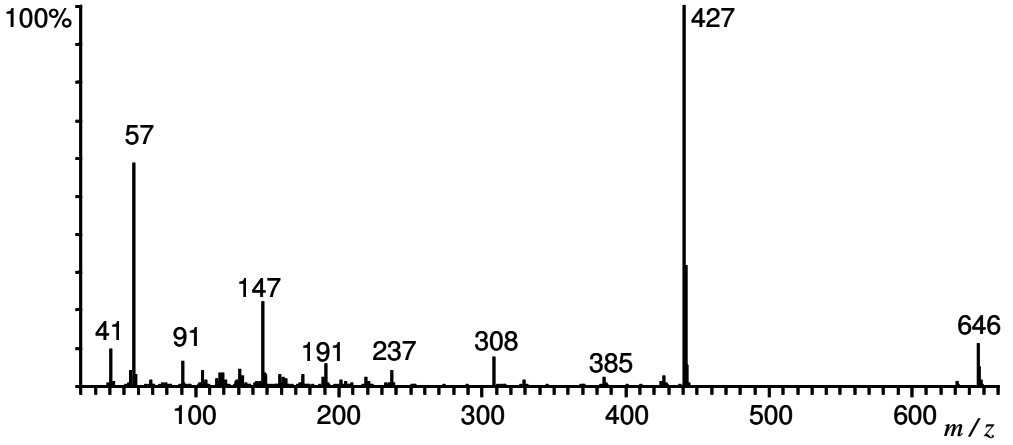


Fig.8.2 Tris-(2,4-di-tertbutylphenyl)-phosphite (**19b**), M646, from 1/Hel12.

8.1.2 Diketopiperazines

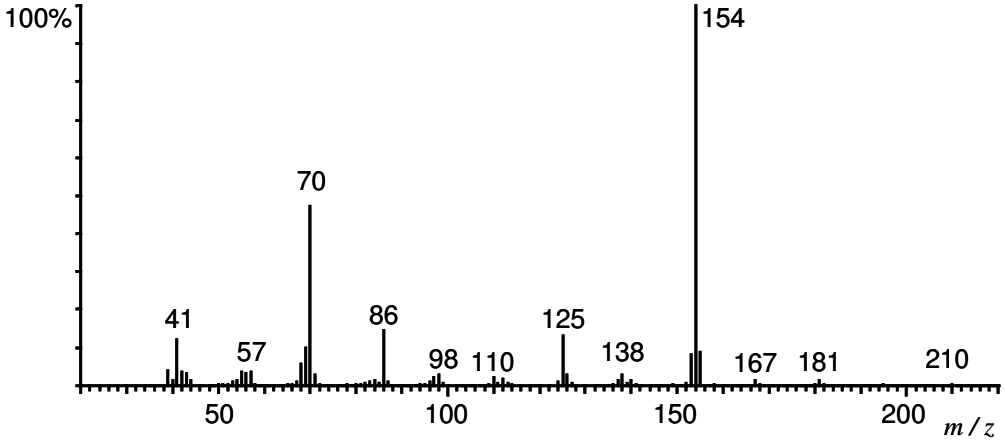


Fig. 8.3 cyc(Ile-Pro) (**20**), M210, from Hel26.

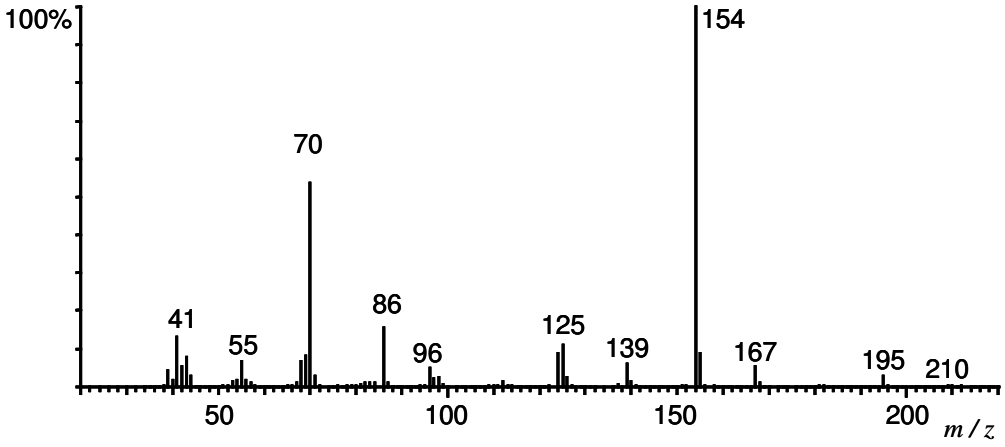


Fig. 8.4 cyc(Leu-Pro), M210, from Hel26.

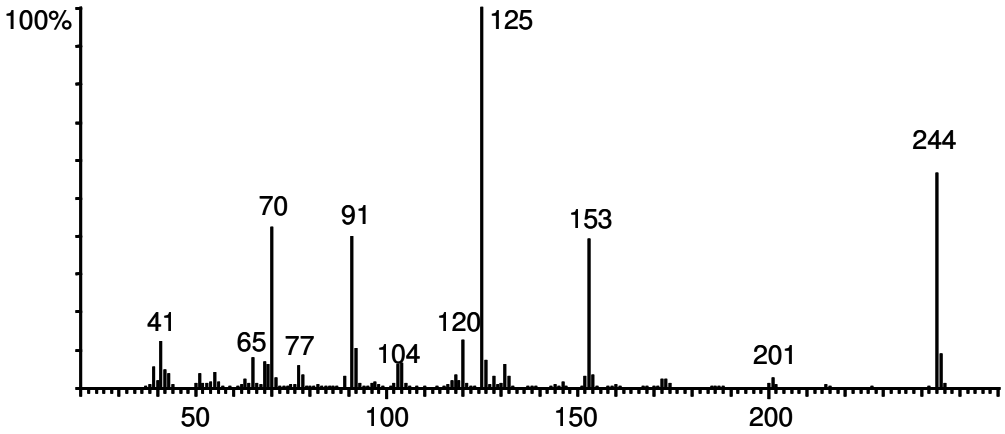


Fig 8.5 cyc(Phe-Pro), M244, from Hel26.

8.1.3 Amides

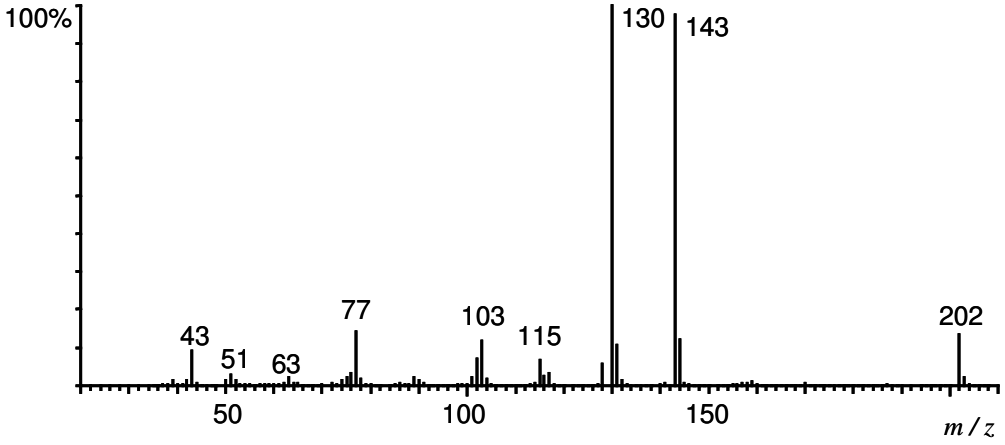


Fig. 8.6 *N*-acetyltryptamine (**23**), M202, from ANT253(RE).

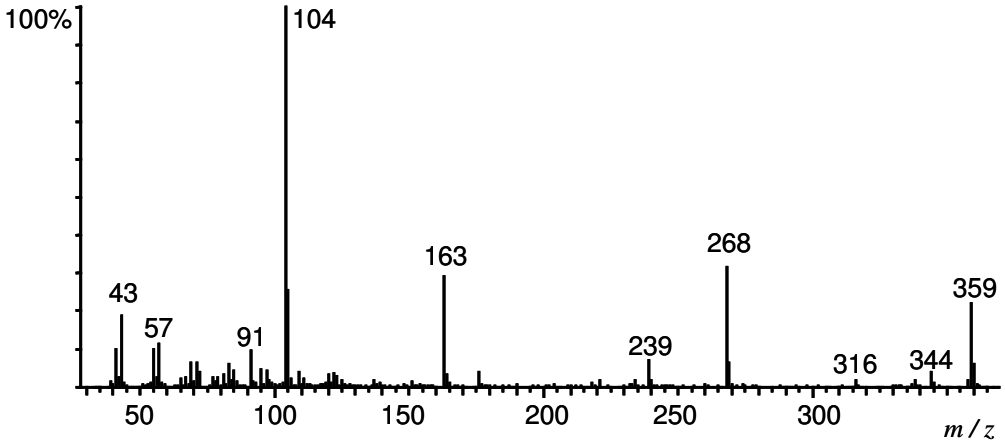


Fig. 8.7 *N*-palmitoyl-2-phenylethylamine (**24**), M359, from Pic006.

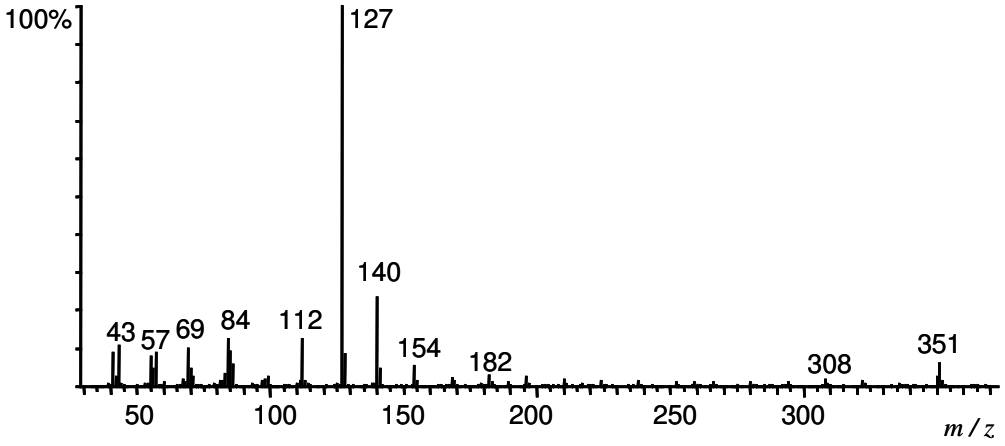


Fig. 8.8 *N*-stearoylpiperidine (**25**), M351, from Hel45.

8.1.4 Hexanetriols

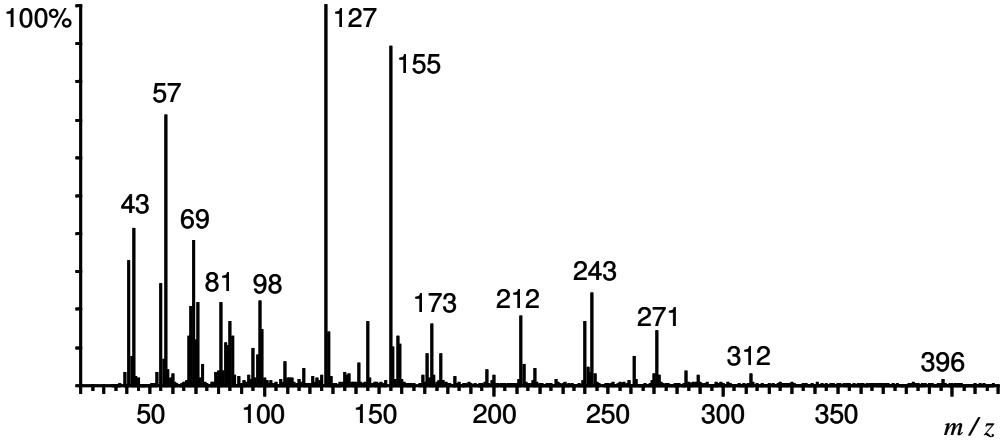


Fig. 8.9 Hexantriol (**30**), M414, *RI* 2801, from 1/IHg,3.

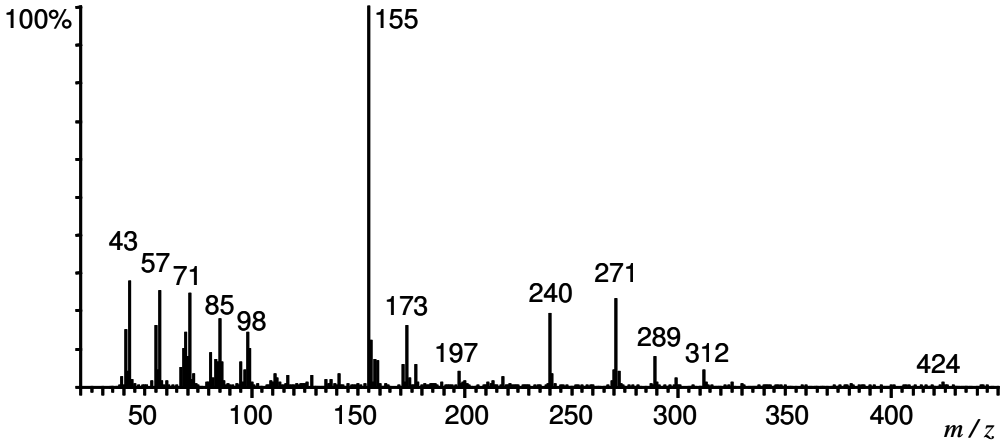


Fig. 8.10 Hexantriol (**31**), M442, *RI* 2999, from 1/IHg,3.

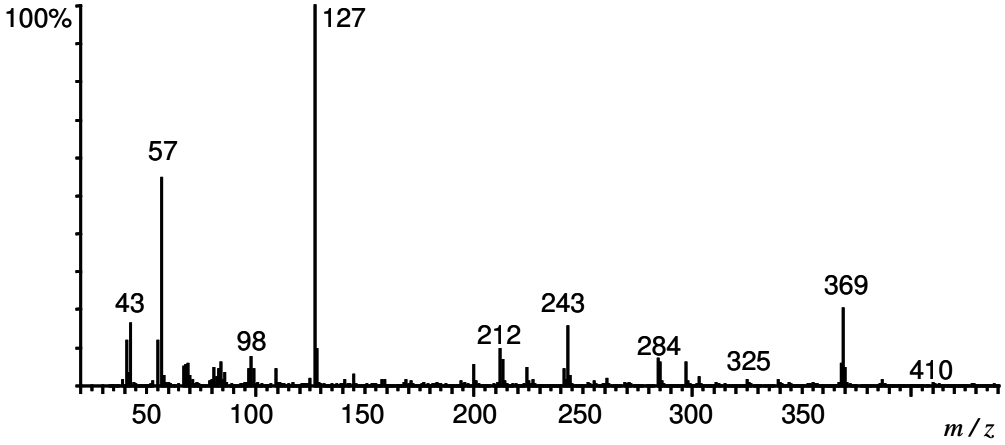


Fig. 8.11 Hexantriol (**32**), M512, *RI* 3149, from 1/IHg,3.

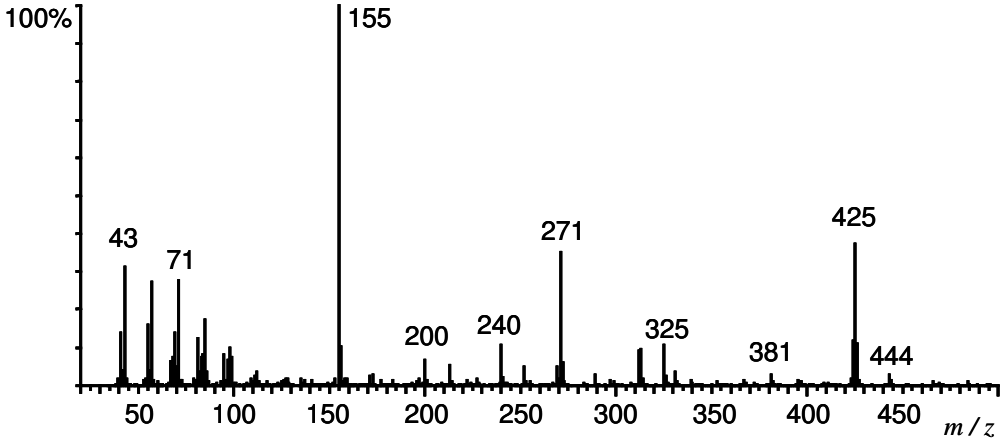


Fig. 8.12 Hexantriol (**35**), M596, *RI* 3709, from 1/IHg,3.

8.1.5 Long-chain ketones and alkenes

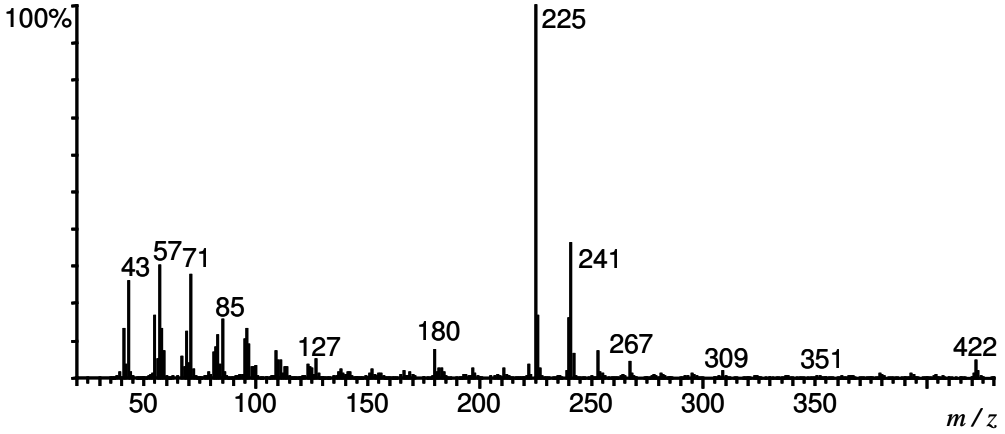


Fig. 8.13 15-nonacosanone (**41**), *RI* 3088, isolated from cabbage.

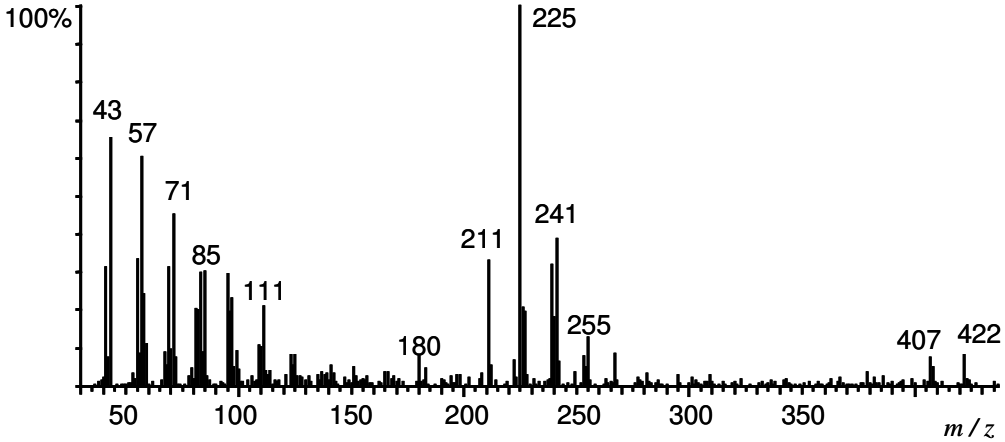


Fig. 8.14 2,26-dimethyl-14-heptacosanone (**40a**), *RI* 3008, from the extract Bio1 (*m/z* 211 and 239 are probably from 2,26-dimethyl-13-heptacosanone).

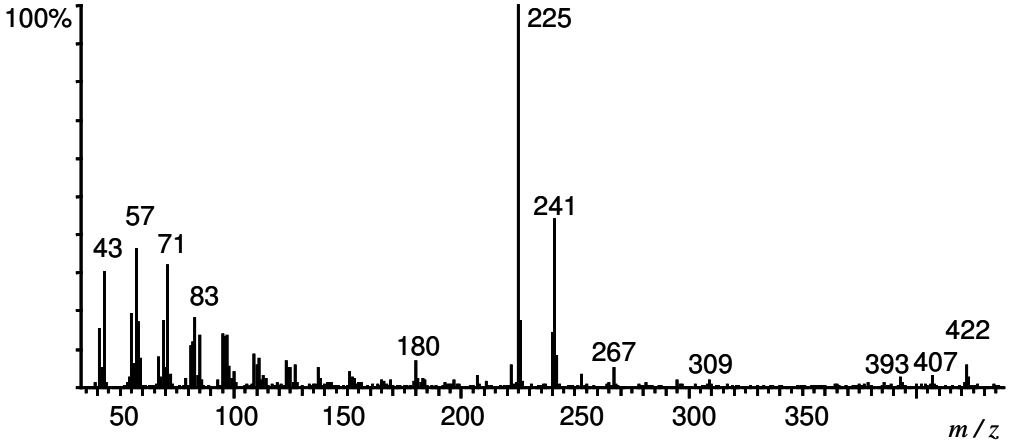


Fig. 8.15 2,25-dimethyl-14-heptacosanone (**40b**), *RI* 3019, from the extract Bio1.

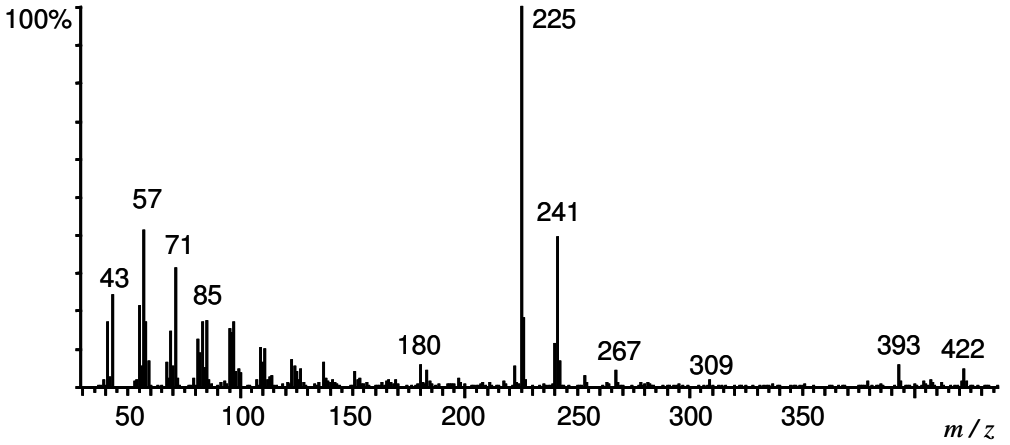


Fig. 8.16 3,25-dimethyl-14-heptacosanone (**40c**), *RI* 3031, from the extract Bio1.

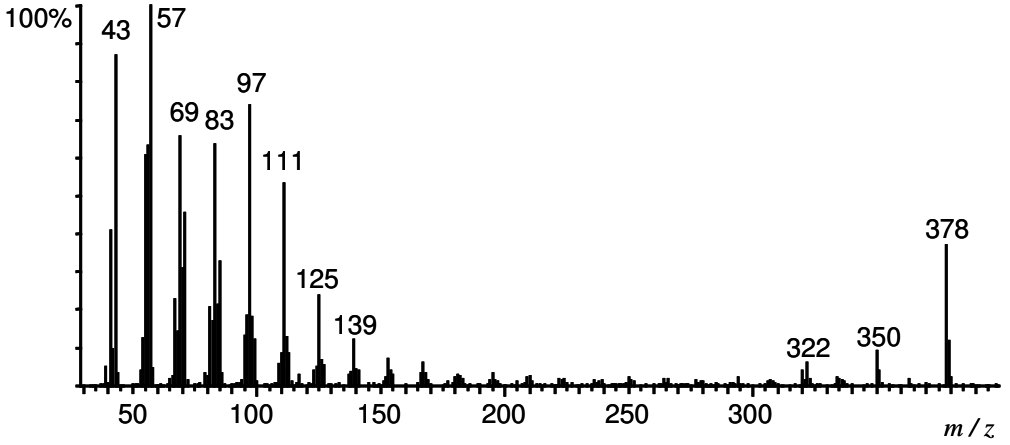


Fig. 8.17 2,24-dimethyl-12-pentacosene (**43a**), *RI* 2585, from the extract Bio1.

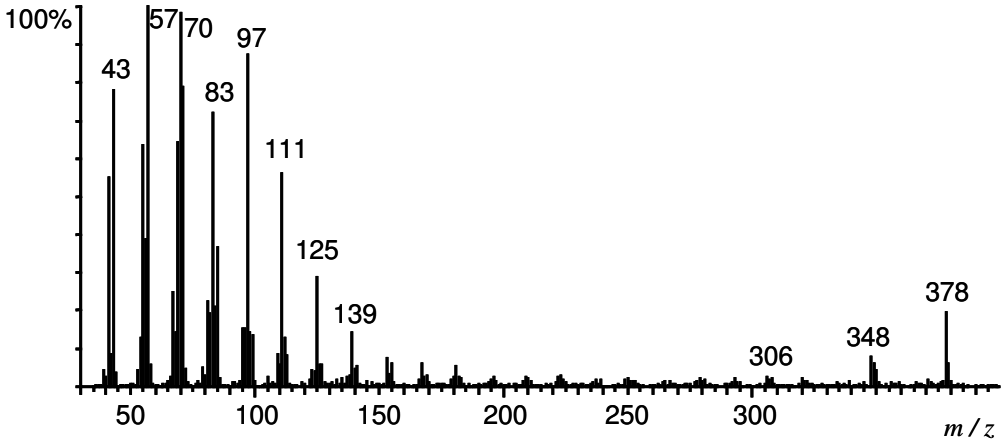


Fig. 8.18 2,23-dimethyl-12-pentacosene (**43b**), *RI* 2595, from the extract Bio1.

8.1.6 Cyclic polysulfides

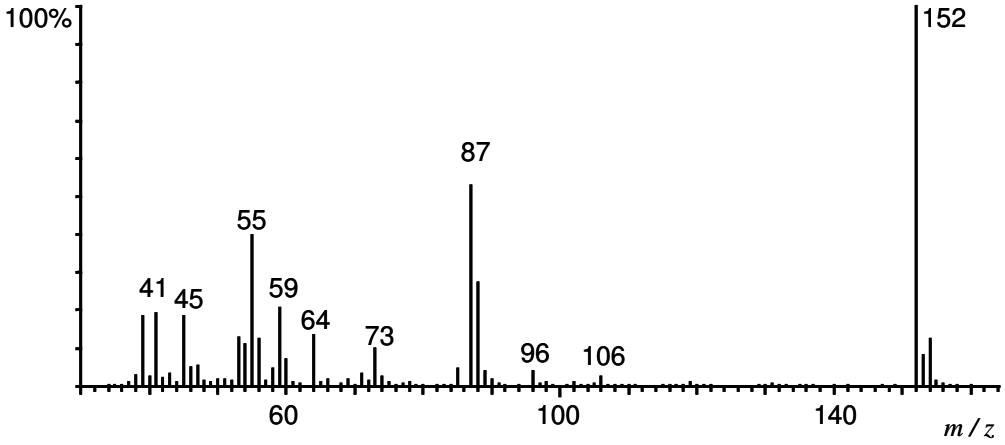


Fig. 8.19 Cyclic polysulfide (**46**), M152, *RI* 1175, from the extract Bio137.

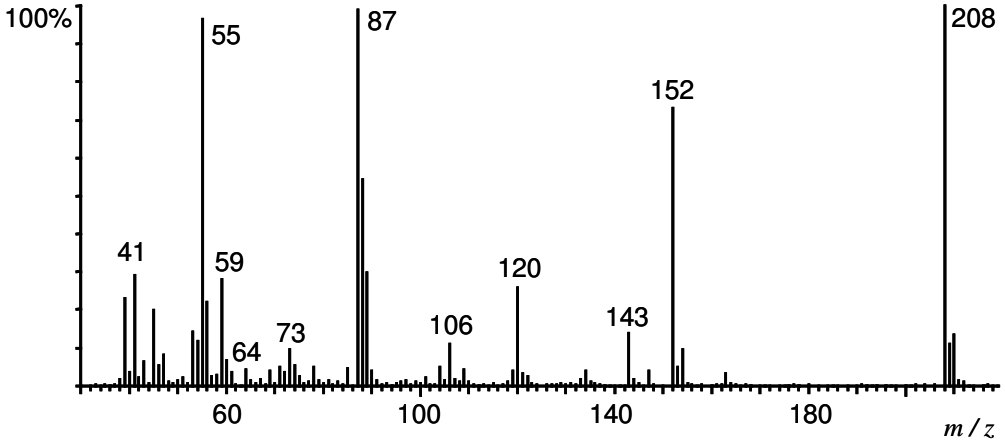


Fig. 8.20 Cyclic polysulfide (**47a**), M208, *RI* 1472, from the extract Bio137.

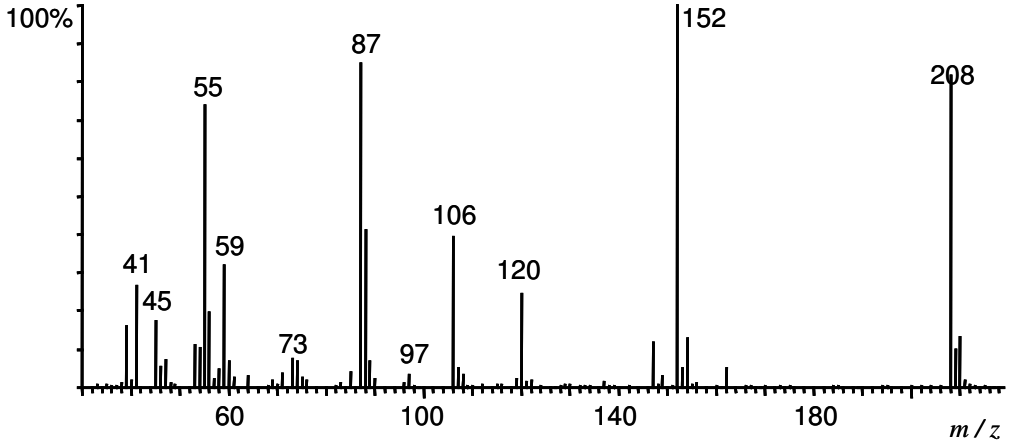


Fig. 8.21 Cyclic polysulfide (**47b**), M208, *RI* 1516, from the extract Bio137.

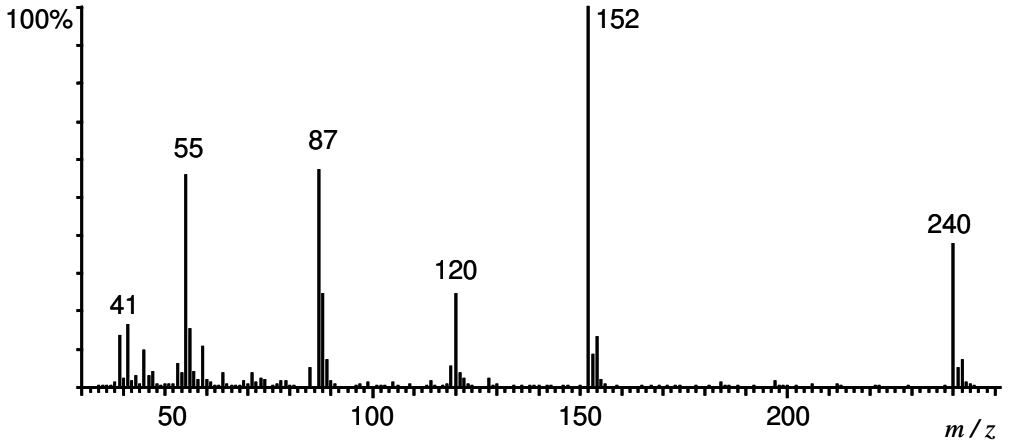


Fig. 8.22 Cyclic polysulfide (**48a**), M240, *RI* 1727, from the extract Bio137.

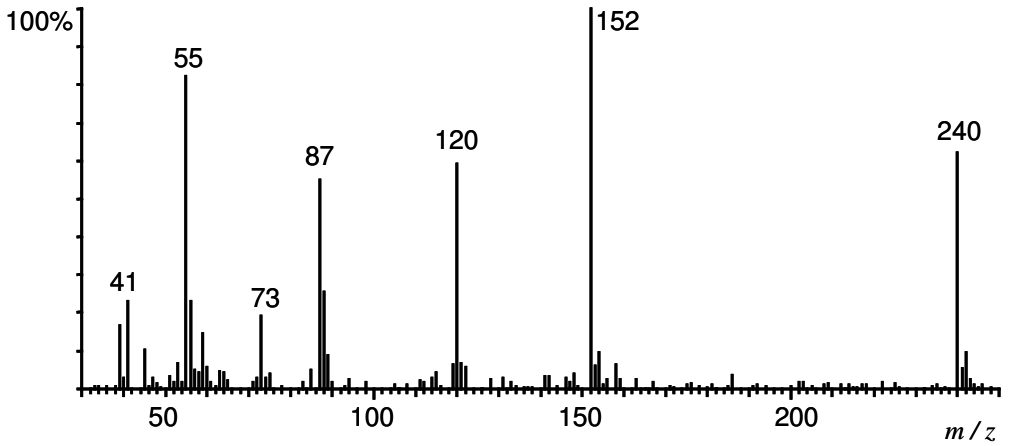


Fig. 8.23 Cyclic polysulfide (**48b**), M240, *RI* 1748, from the extract Bio137.

8.1.7 Isomeric lactones

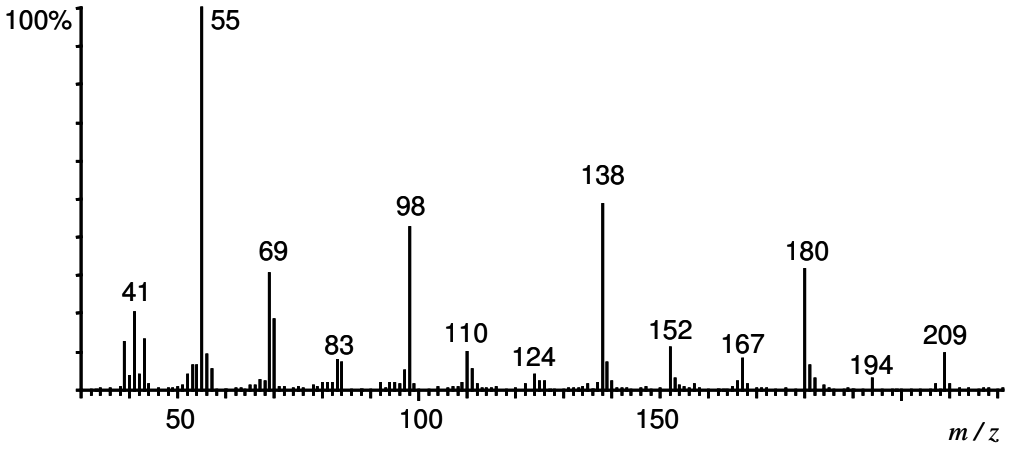


Fig. 8.24 Lactone (**50a**), M209, *RI* 1442, from the extract Hel59 5L.

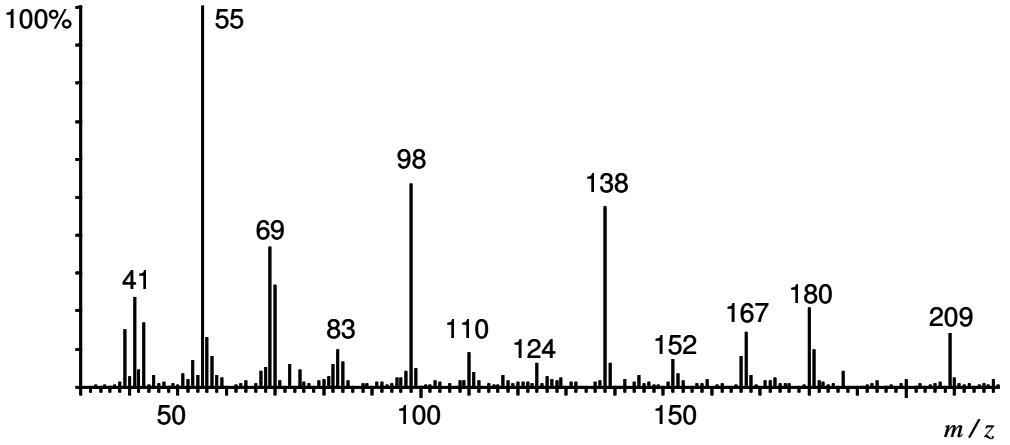


Fig. 8.25 Lactone (**50b**), M209, *RI* 1462, from the extract Hel59 5L.

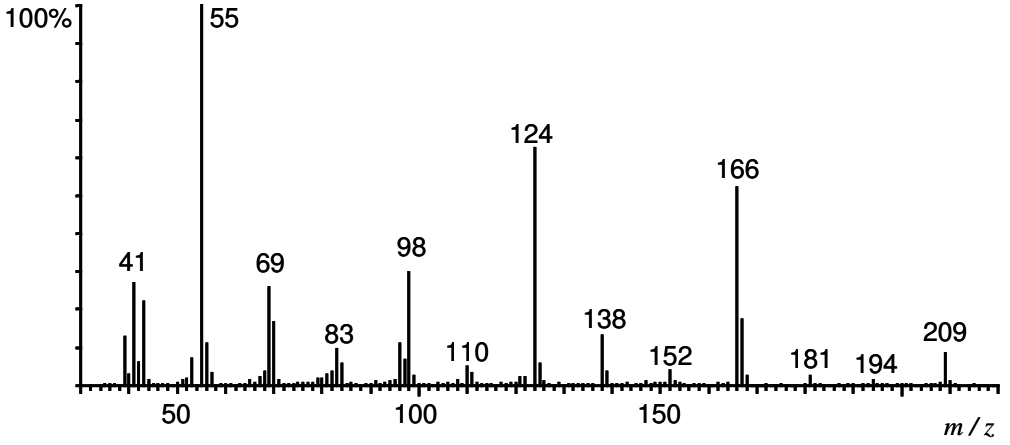


Fig. 8.26 Lactone (**49a**), M209, *RI* 1454, from the extract Hel59 5L.

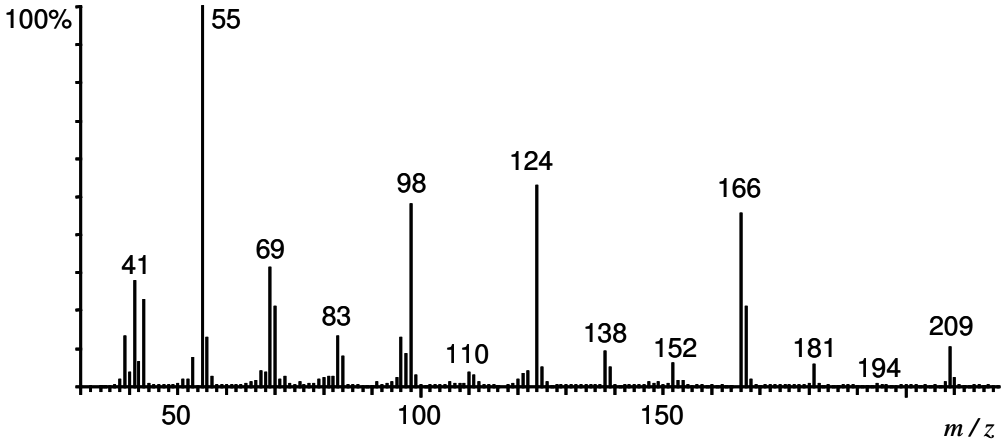


Fig. 8.27 Lactone (**49b**), M209, *RI* 1468, from the extract Hel59 5L.

8.2 GC-IR spectra

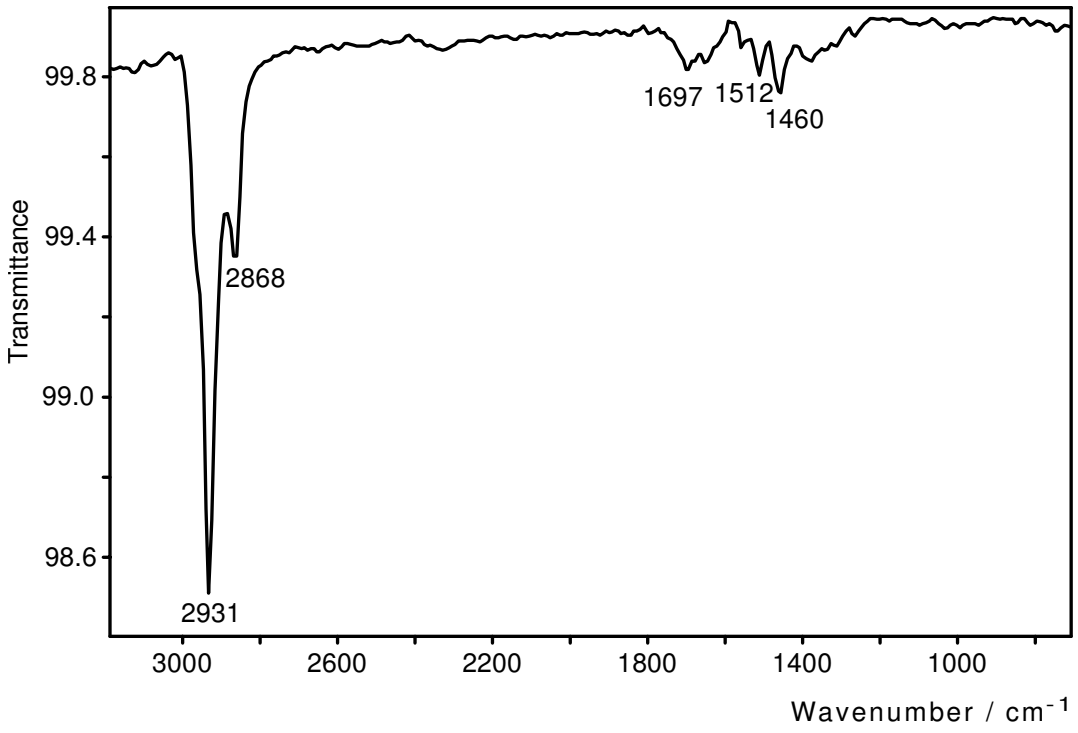


Fig. 8.28 Gas-phase IR spectrum of 3,25-dimethyl-13-heptacosene (**45c**), *RI* 2803, from Pic006.

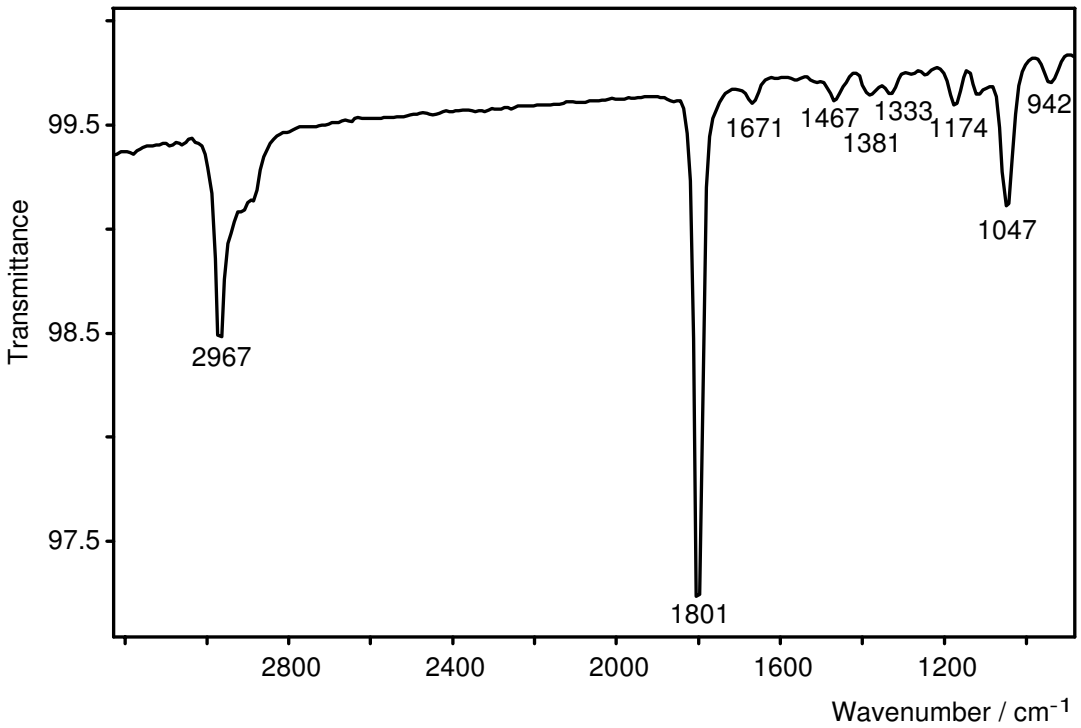


Fig. 8.29 Gas-phase IR spectrum of lactone **49a**, *RI* 1454, from 1/AMP,3.

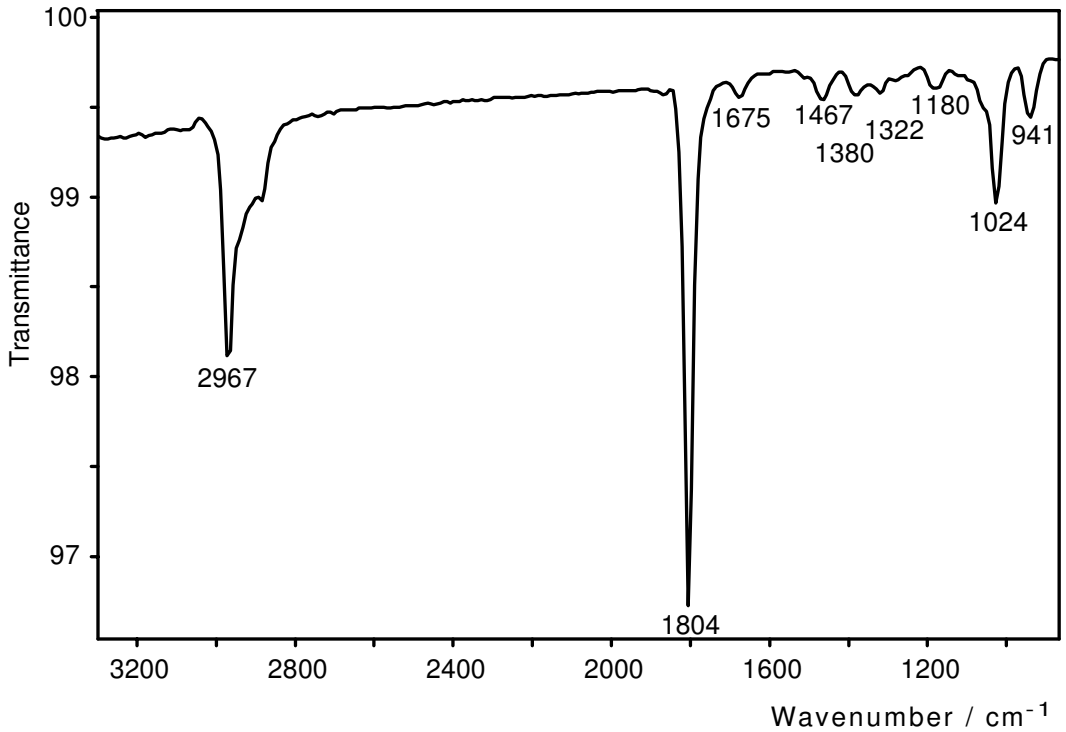


Fig. 8.30 Gas-phase IR spectrum of lactone **49b**, *RI* 1468, from 1/AMP,3.

8.3 Total ion chromatograms (TICs)

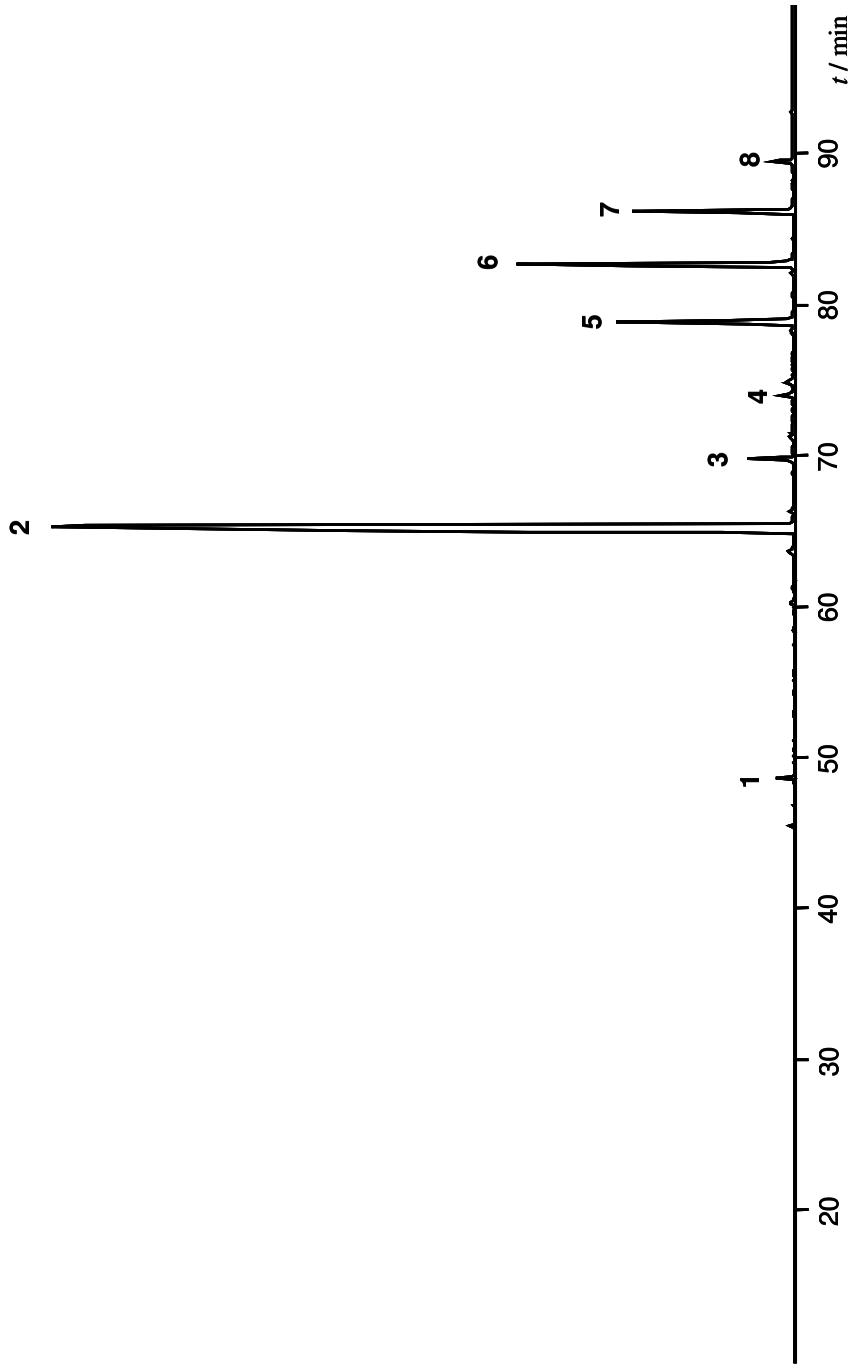


Fig. 8.31 TIC of IHg,3 (*Cytophaga* sp.); (**1**) dibutylphthalate, (**2**) bis-(2-ethylhexyl)-phthalate, (**3**) hexanetriol **30**-TMS, (**4**) hexanetriol **31**-TMS, (**5**) hexanetriol **32**, (**6**) hexanetriol **33**, (**7**) hexanetriol **34**, (**8**) hexanetriol **35**.

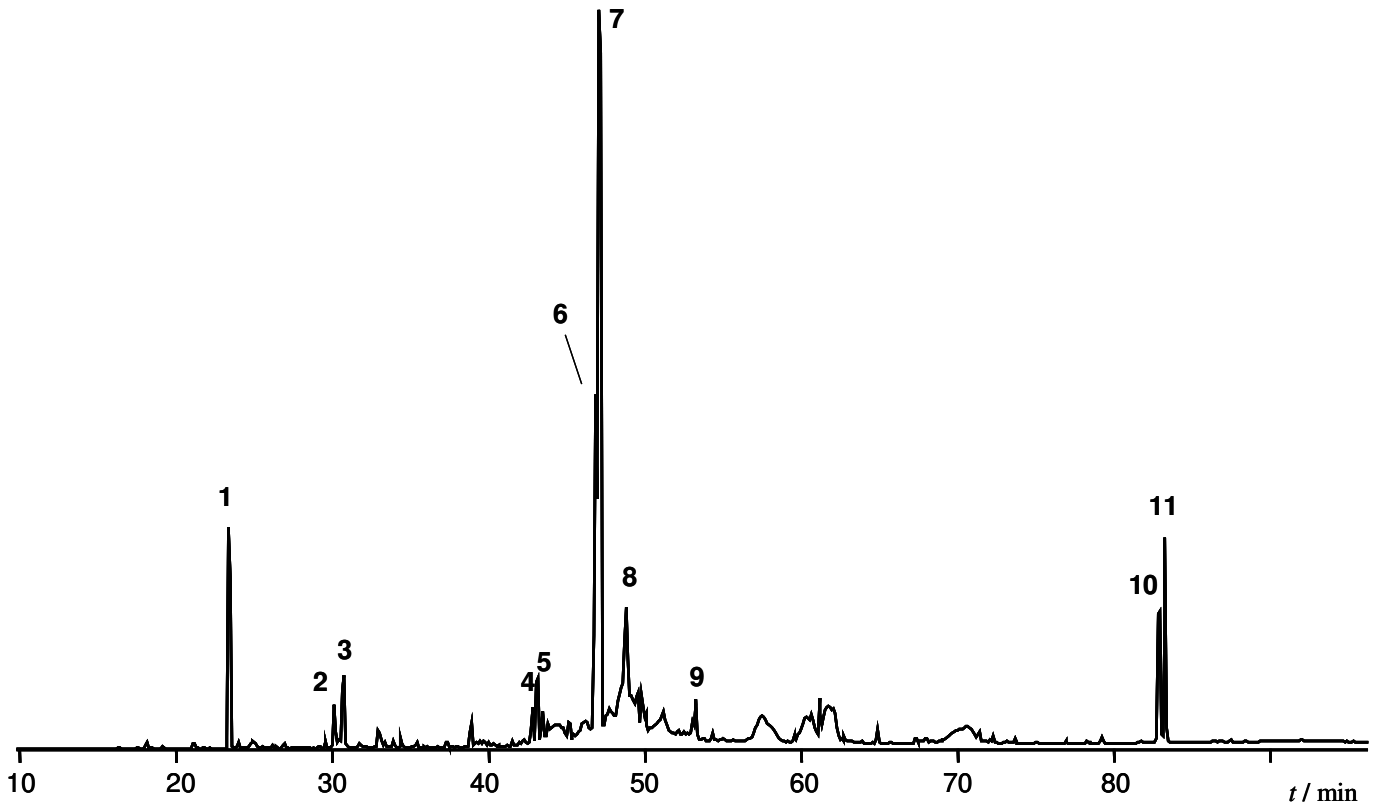


Fig. 8.32 TIC of 1/AMP,3 (*Cytophaga* sp.); (**1**) 2-phenylacetic acid TMS ester, **22**-TMS (**2**) lactone **49a**, (**3**) lactone **49b**, (**4**) methyl 13-methylmyristate, (**5**) methyl 12-methylmyristate, (**6**) 13-methylmyristic acid TMS-ester, (**7**) 12-methylmyristic acid TMS ester, (**8**) cyc(Pro-Pro), (**9**) 14-methylpalmitic acid TMS ester, (**10**) C35 terpene **26**, (11) C35 terpene **27**.

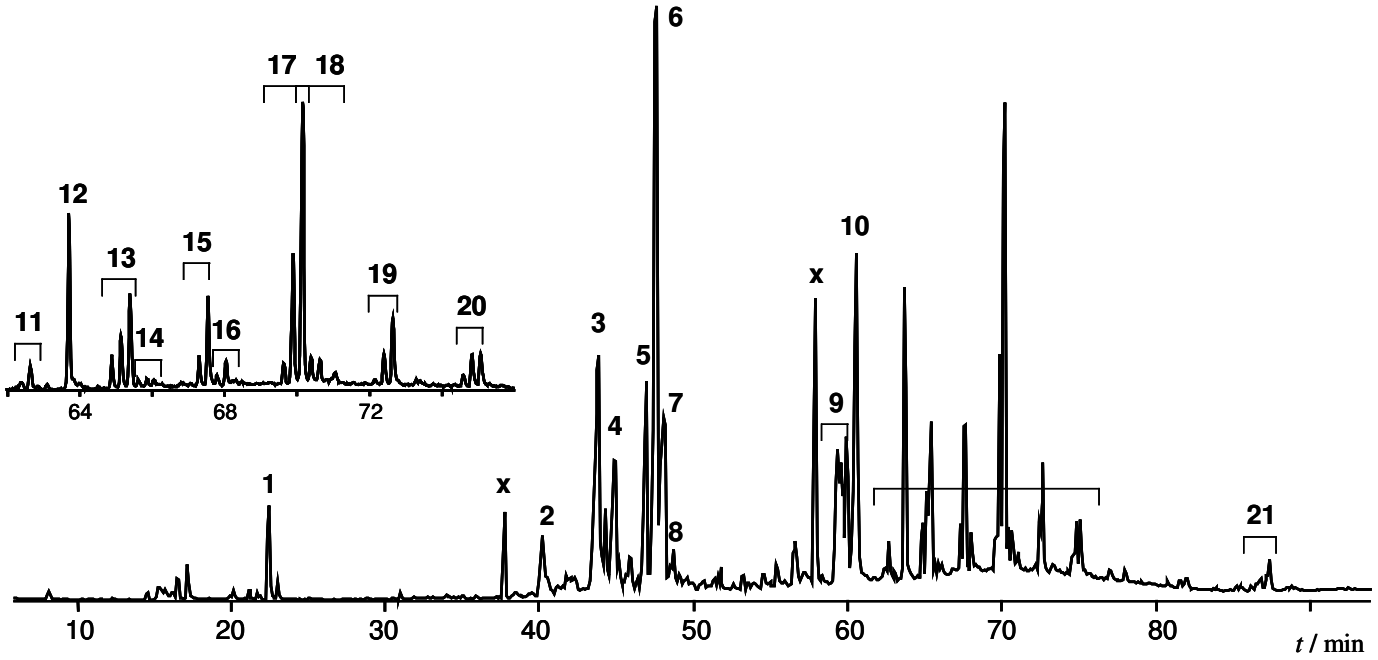


Fig. 8.33 TIC of Bio1 (*Cytophaga* sp.); (**1**) 2-phenylacetic acid TMS ester, **22**-TMS, (**2**) cyc(Ala-Pro), ethyl myristate in trace, (**3**) cyc(Pro-Val) 1., ethyl pentadecanoate in trace, (**4**) cyc(Pro-Val) 2., (**5**) cyc(Ile-Pro) 1., (**6**) dibuthyl-phthalate, cyc(Leu-Pro) 1., (**7**) cyc(Leu-Pro) 2., cyc(Ile-Pro) 2., (**8**) ethyl palmitate, (**9**) cyc(Phe-Pro) 1., 9-octadecen-amide (2 isomers), bis-(2-ethylhexyl)-adipate, (**10**) cyc(Phe-Pro) 2., (**11**) alkenes **42a**, **b**, (**12**) bis-(2-ethylhexyl)-phthalate, (**13**) alkenes **43a**, **b**, **c**, (**14**) ketones **36b**, **c**, (**15**) alkenes **44a**, **b**, (**16**) ketones **37a**, **b**, (**17**) alkenes **45a**, **b**, **c**, (**18**) ketones **38a**, **b**, **c**, (**19**) ketones **39b**, **c**, (**20**) ketones **40a**, **b**, **c**, (**21**) glycerin esters. For compounds **36**-**40** and **42**-**45** see Table 3.12.

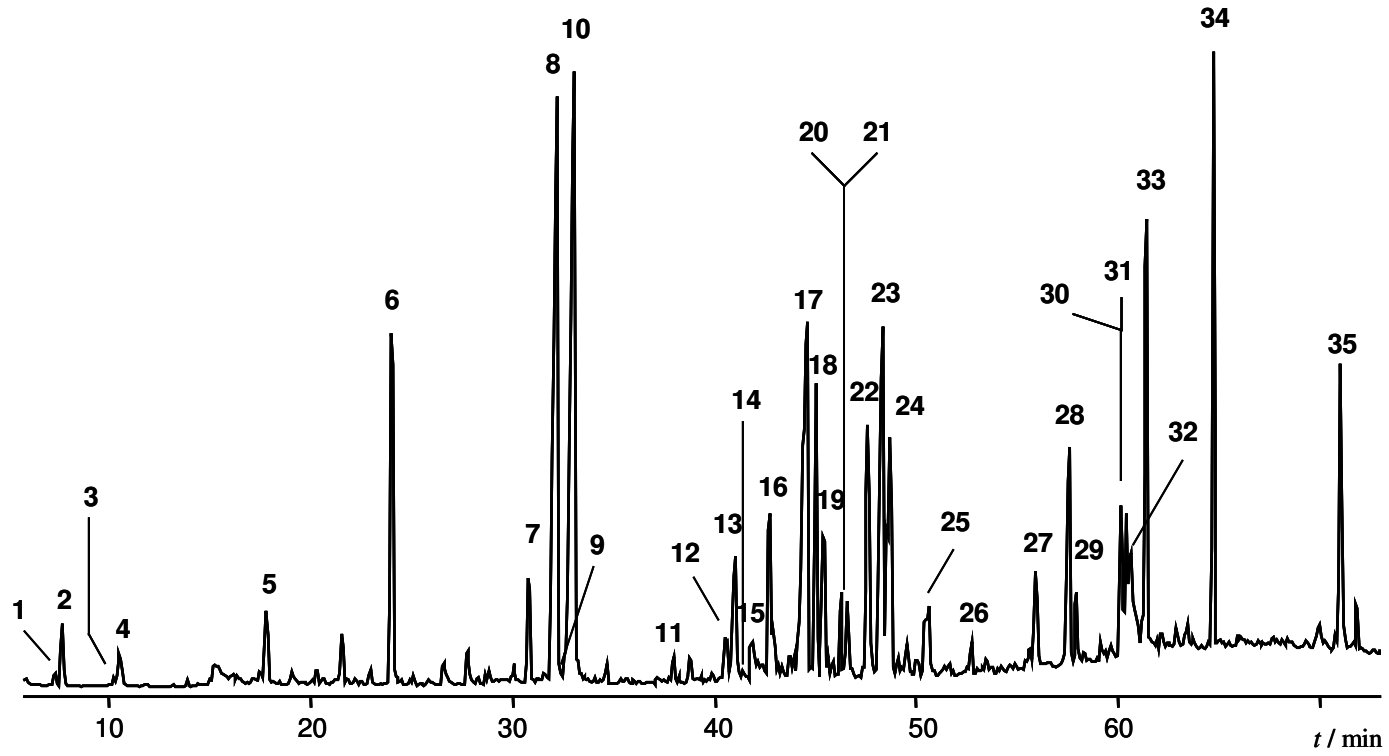


Fig. 8.34 TIC of Bio137 (*Cytophaga* sp.); (**1**) 2-methylbutanoic acid TMS ester, (**2**) 3-methylbutanoic acid TMS ester, (**3**) 2-methylbutenoic acid TMS ester, (**4**) 3-methylbutenoic acid TMS ester, (**5**) cyclic polysulfide **46**, (**6**) (2-phenyl-ethyl)-amine, (**7**), cyclic polysulfide **47a**, (**8**) *N*-(2-phenylethyl)-formamide, (**9**) cyclic polysulfide **47b**, (**10**) *N*-acetyl-(2-phenylethyl)-amine, **16a**, (**11**) 13-methyl-3-tetradecanone, (**12**) cyclic polysulfide 48a, (**13**) cyc(Ala-Pro), (**14**) cyclic polysulfide 48b, (**15**) 2-hexadecanone, (**16**) methyl 13-methylmyristate, (**17**) cyc(Pro-Val) 1., (**18**) ethyl 13-methylmyristate, (**19**) cyc(Pro-Val) 2., (**20**) ethyl pentadecanoate, (**21**) methyl palmitate, (**22**) cyc(Ile-Pro) 1., (**23**) cyc(Leu-Pro) 1., (**24**) cyc(Leu-Pro) 2., cyc(Ile-Pro) 2., ethyl hexadecenoate, (**25**) ethyl heptadecenoate, (**26**) methyl octadecenoate, (**27**) N-acetyltryptamine, 23, (**28**) cyc(Met-Pro) 1., (**29**) cyc(Met-Pro) 2., (**30**) cyc(Phe-Pro) 1., (**31,32**) 9-octadecenamide, (**33**) cyc(Phe-Pro) 2., (**34**) bis(2-ethylhexyl)-phthalate, (**35**) bis(2-ethylhexyl)-adipate.

8.4 Dendograms

Marked sample names

Dendogram 8.1-8.12: gray: group of similar samples or single outliers (alkane standard) that do not belong to the data set,

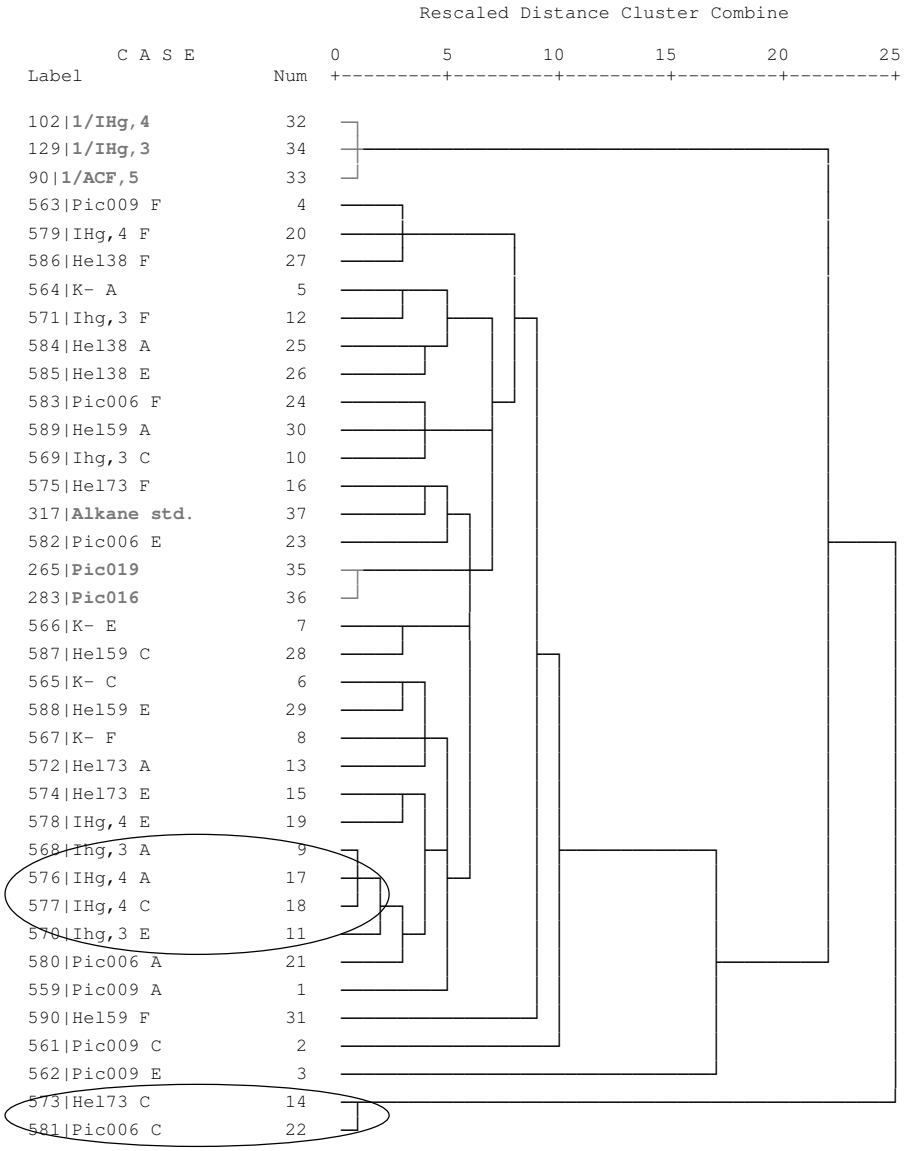
Dendogram 8.3-8.12: underlined: sample analyzed by GC-MS,
italics: injection and/or derivatization control sample,
*: sample fermented more times,
Ros: *Roseobacter* sp.,

Dendogram 8.11, 8.12: **bold**: sample contains the C35 terpenes **26** and **27**.

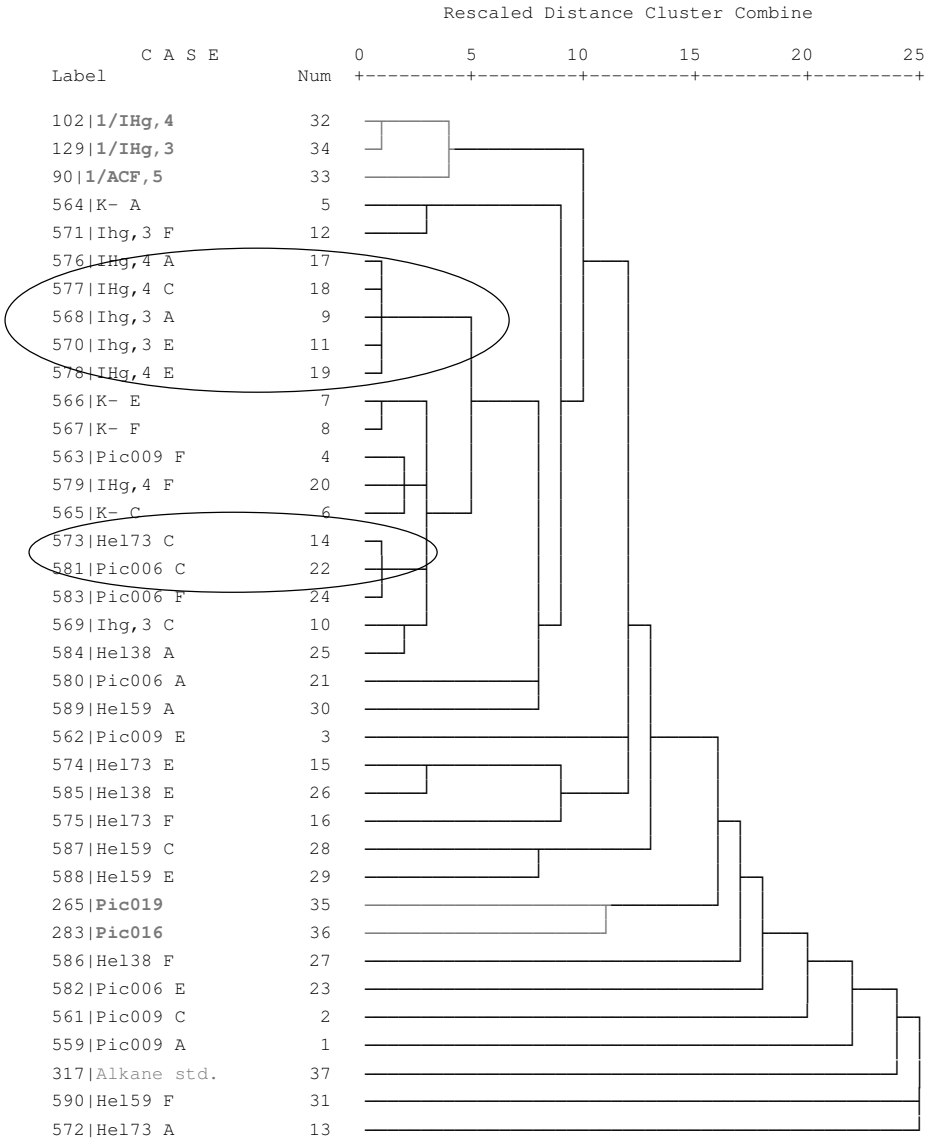
Substance codes

Examples	C16:0-(O)OH:	palmitic acid,
	3-OH-15-Me-C16:0-Me:	methyl 3-hydroxy-15-methylpalmitate,
	9-C18:1-NH_2:	9-octadecenamide,
	FA-Et:	fatty acid ethyl ester.

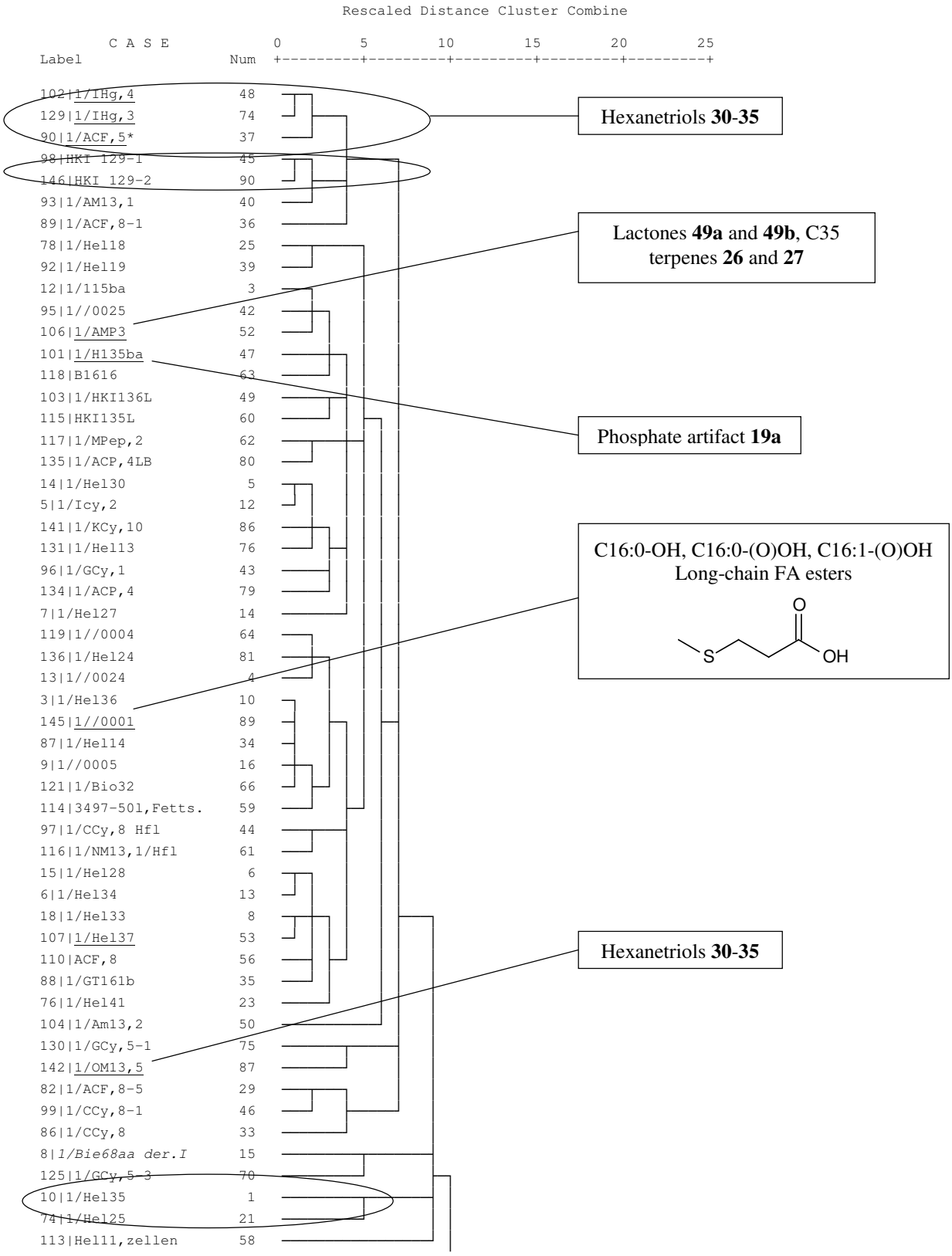
Dendogram 8.1 Parallel fermentation data set; HCA, average linkage, quadratic Euclidean distance.



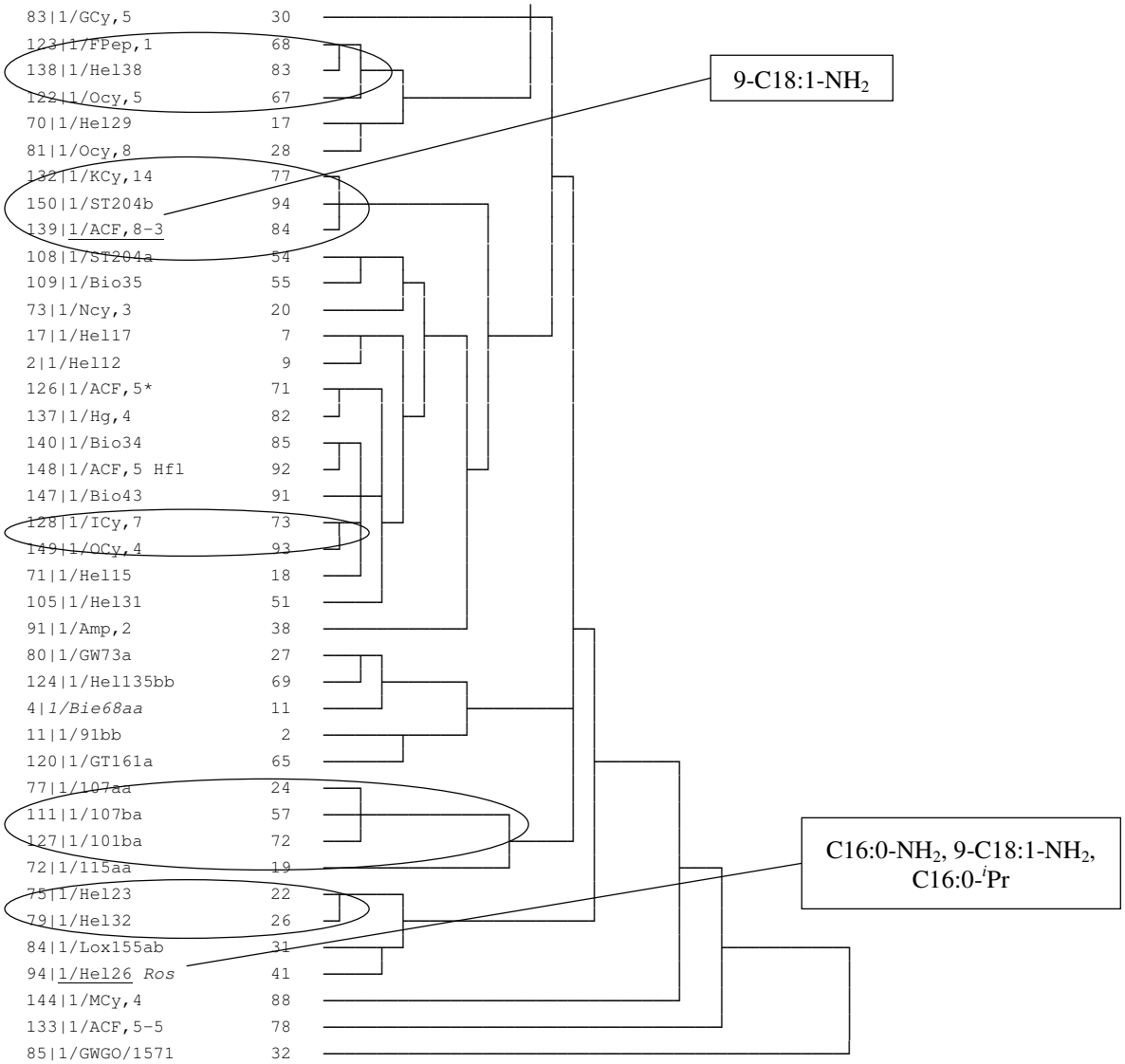
Dendogram 8.2 Parallel fermentation data set; PCA/HCA, 75% of the variance extracted, average linkage, quadratic Euclidean distance.



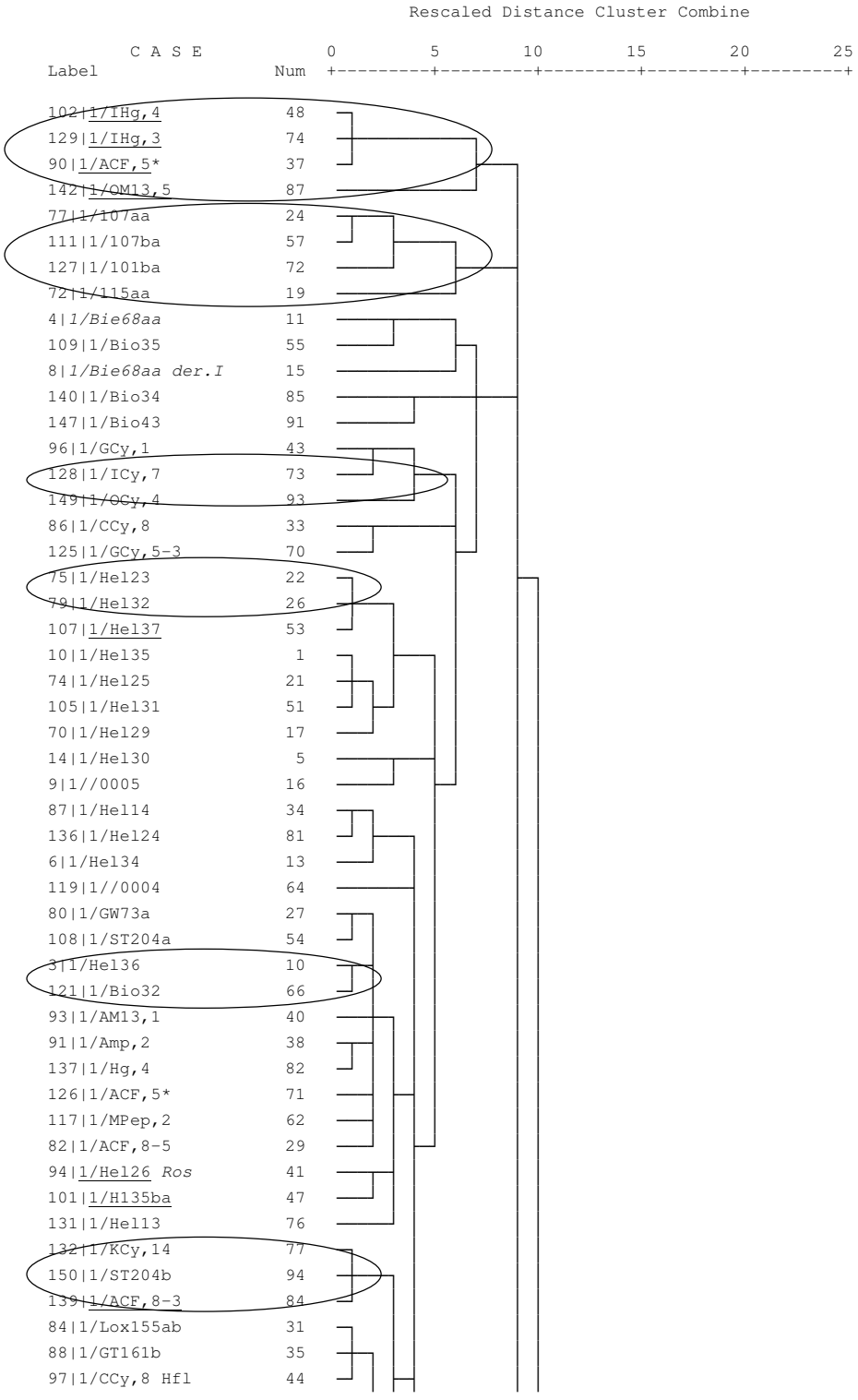
Dendogram 8.3 Subset STM2; HCA, Average linkage, quadratic Euclidean distance



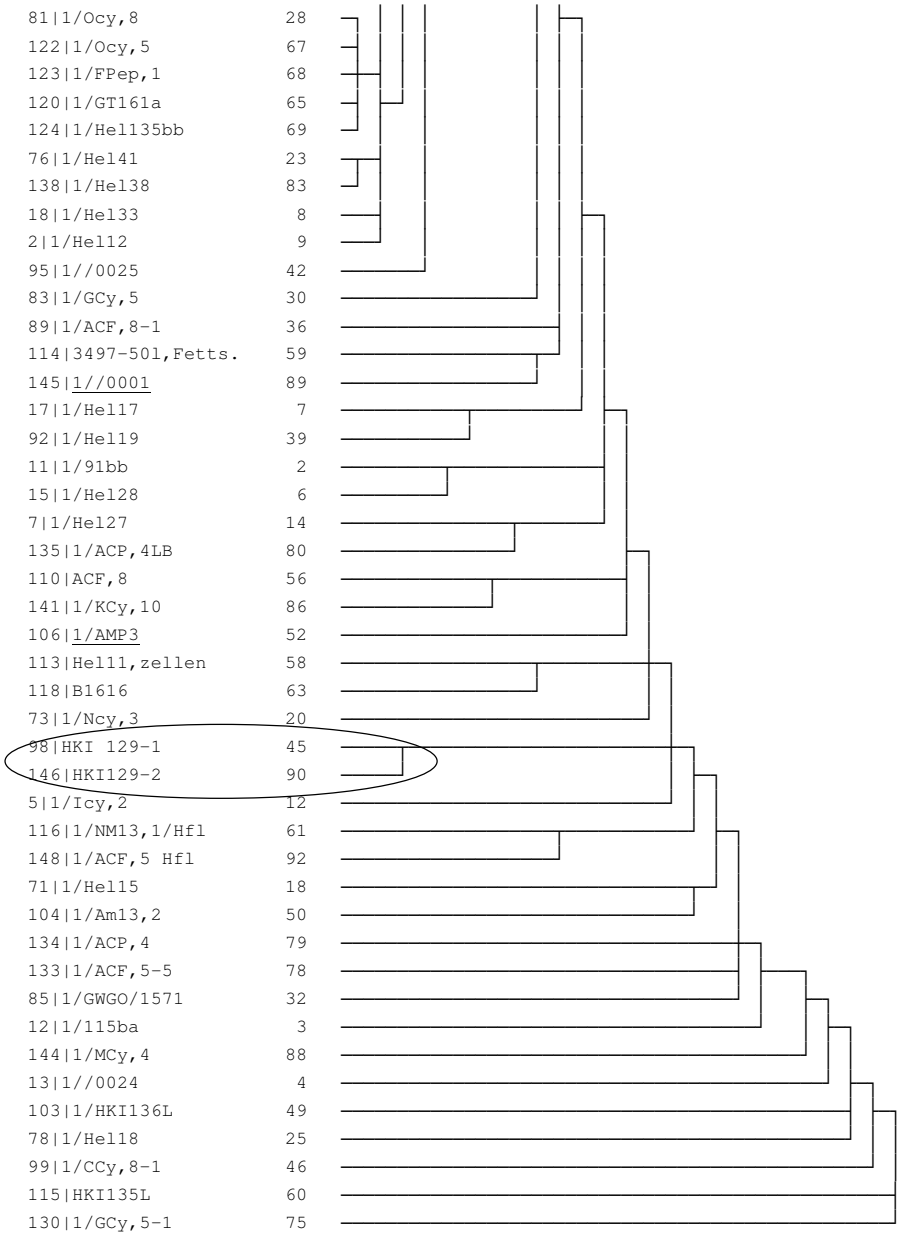
Dendogram 8.3 Continued.



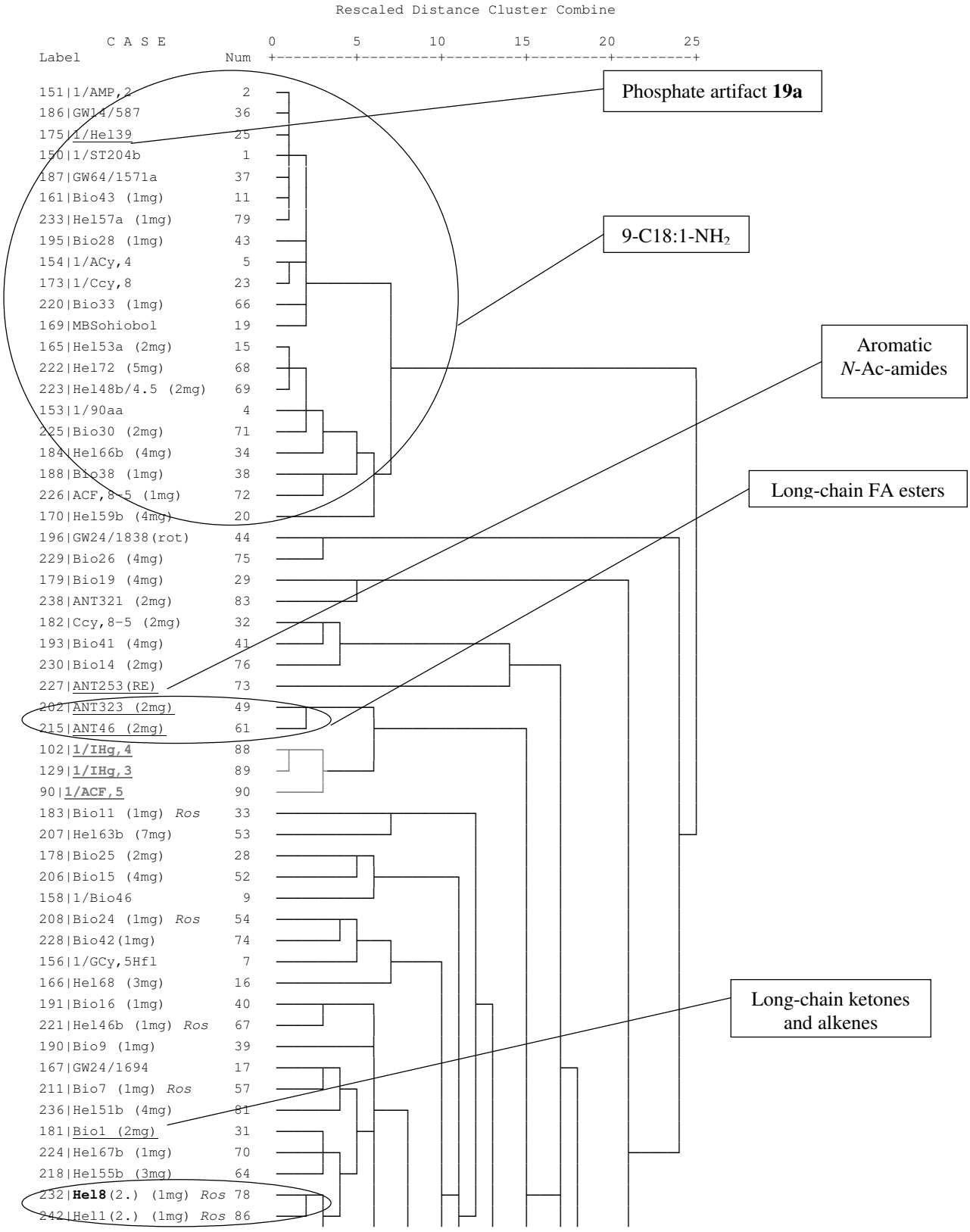
Dendogram 8.4 Subset STM2; PCA/HCA, 79% of the variance extracted, average linkage, quadratic Euclidean distance.



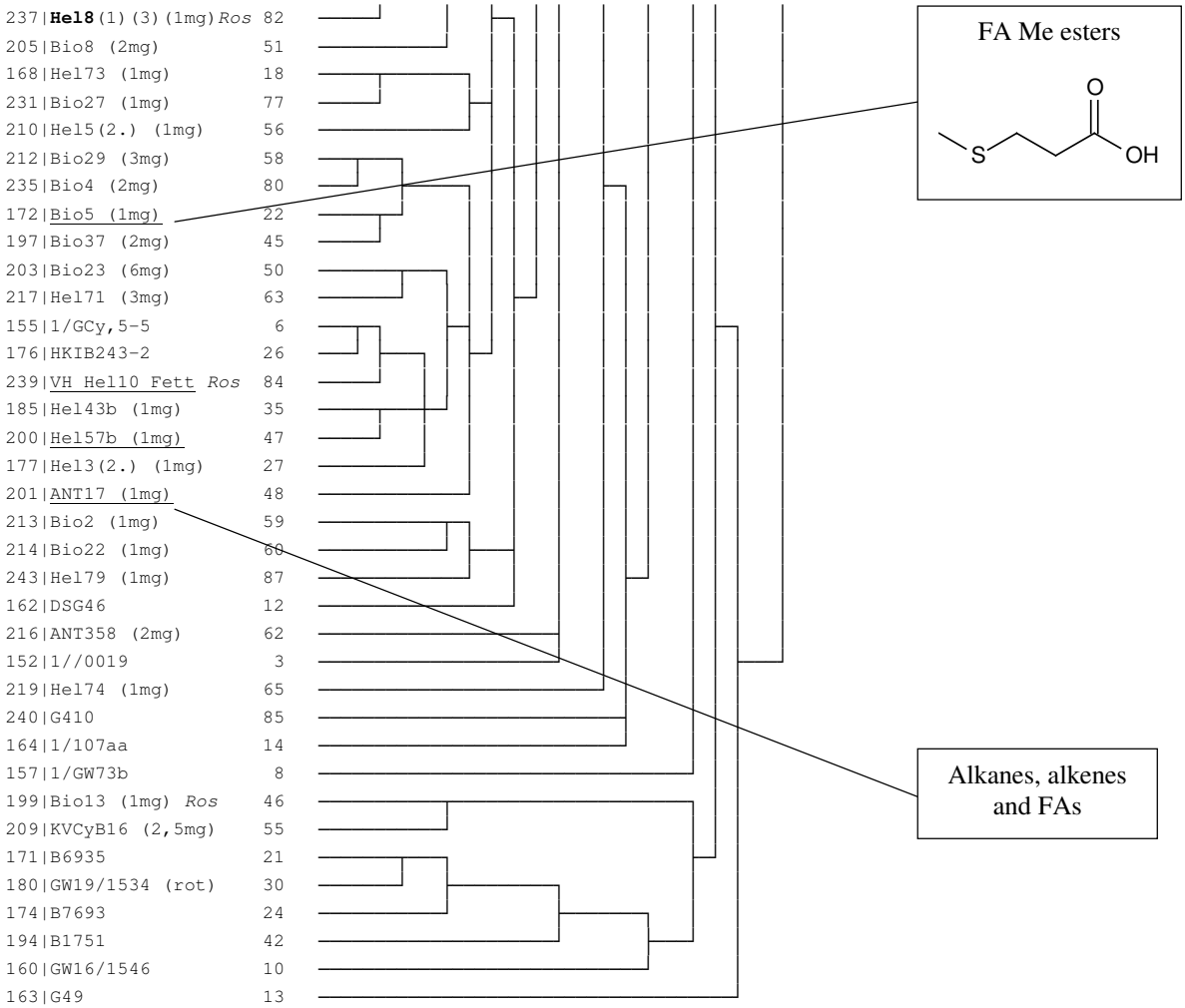
Dendogram 8.4 Continued.



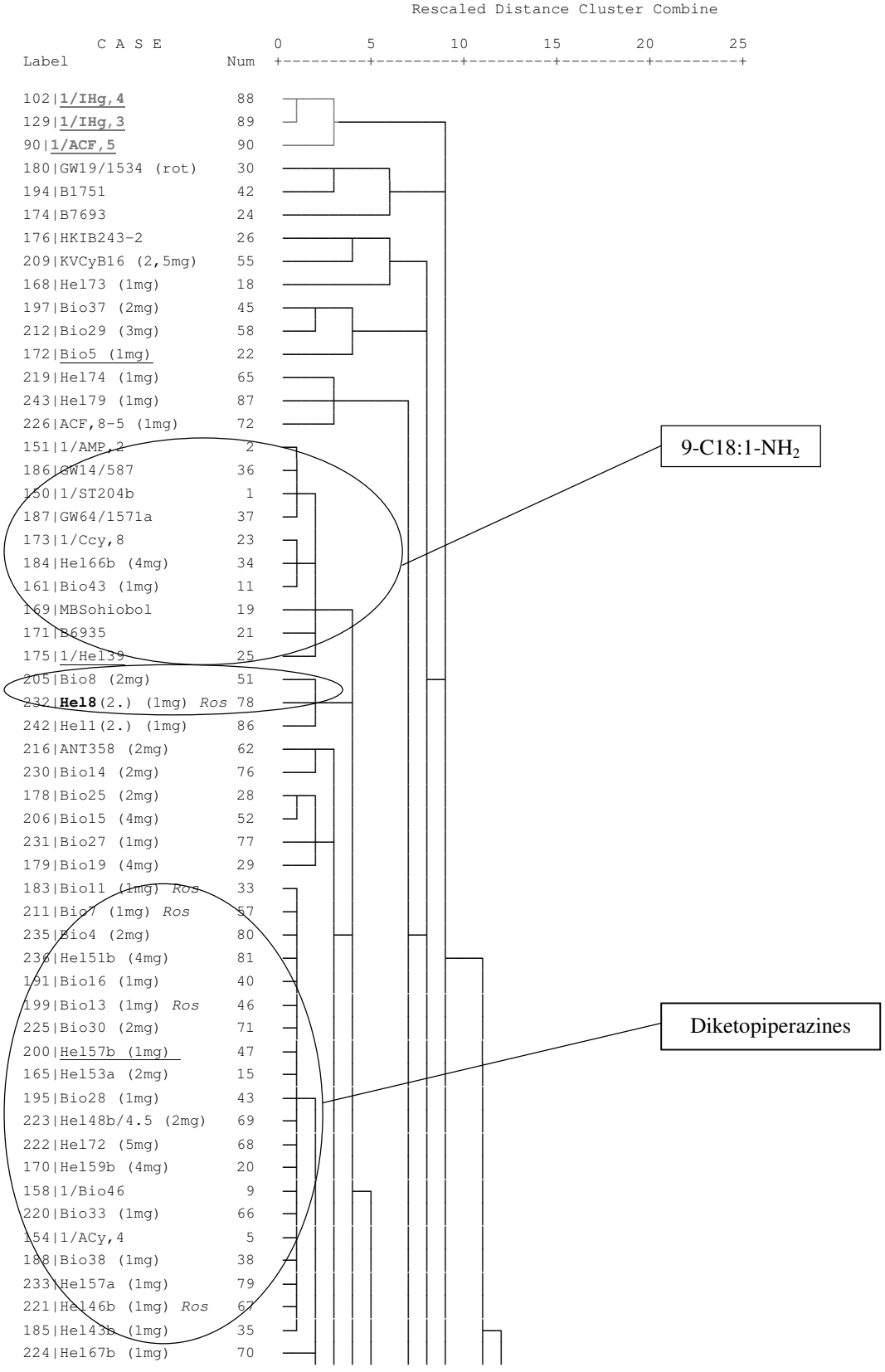
Dendogram 8.5 Subset STM3; HCA, Average linkage, quadratic Euclidean distance.



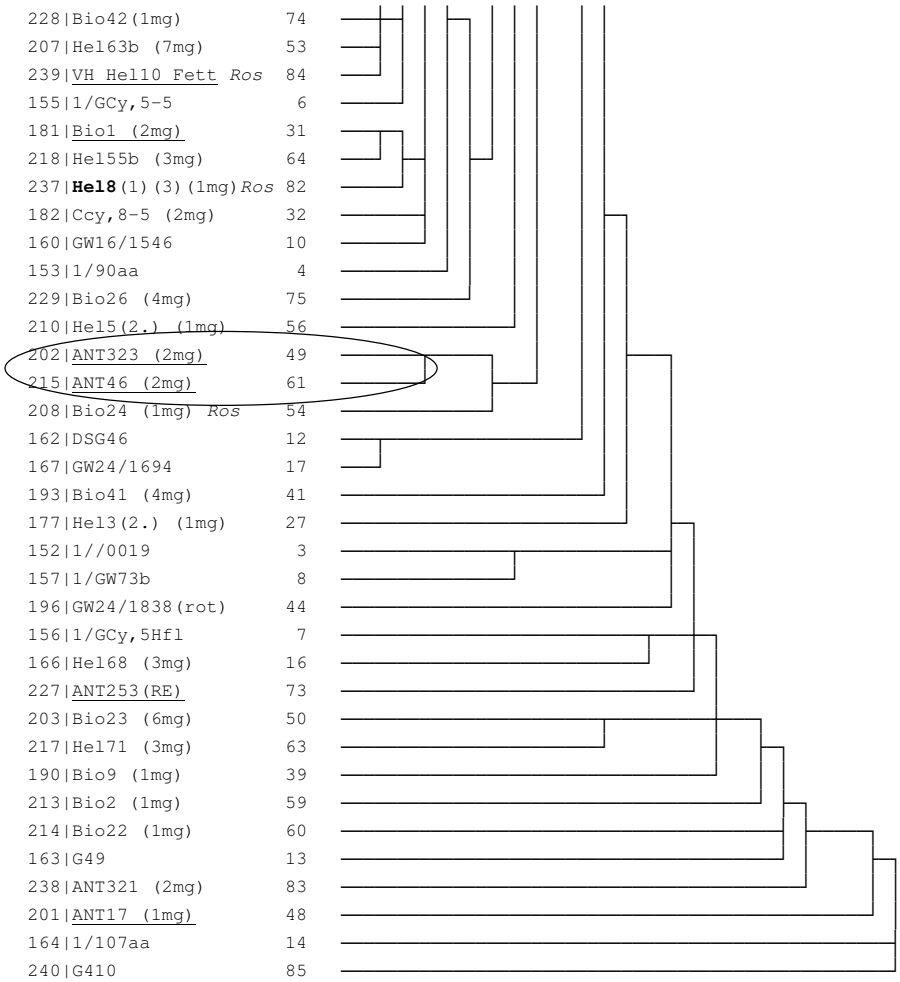
Dendogram 8.5 Continued.



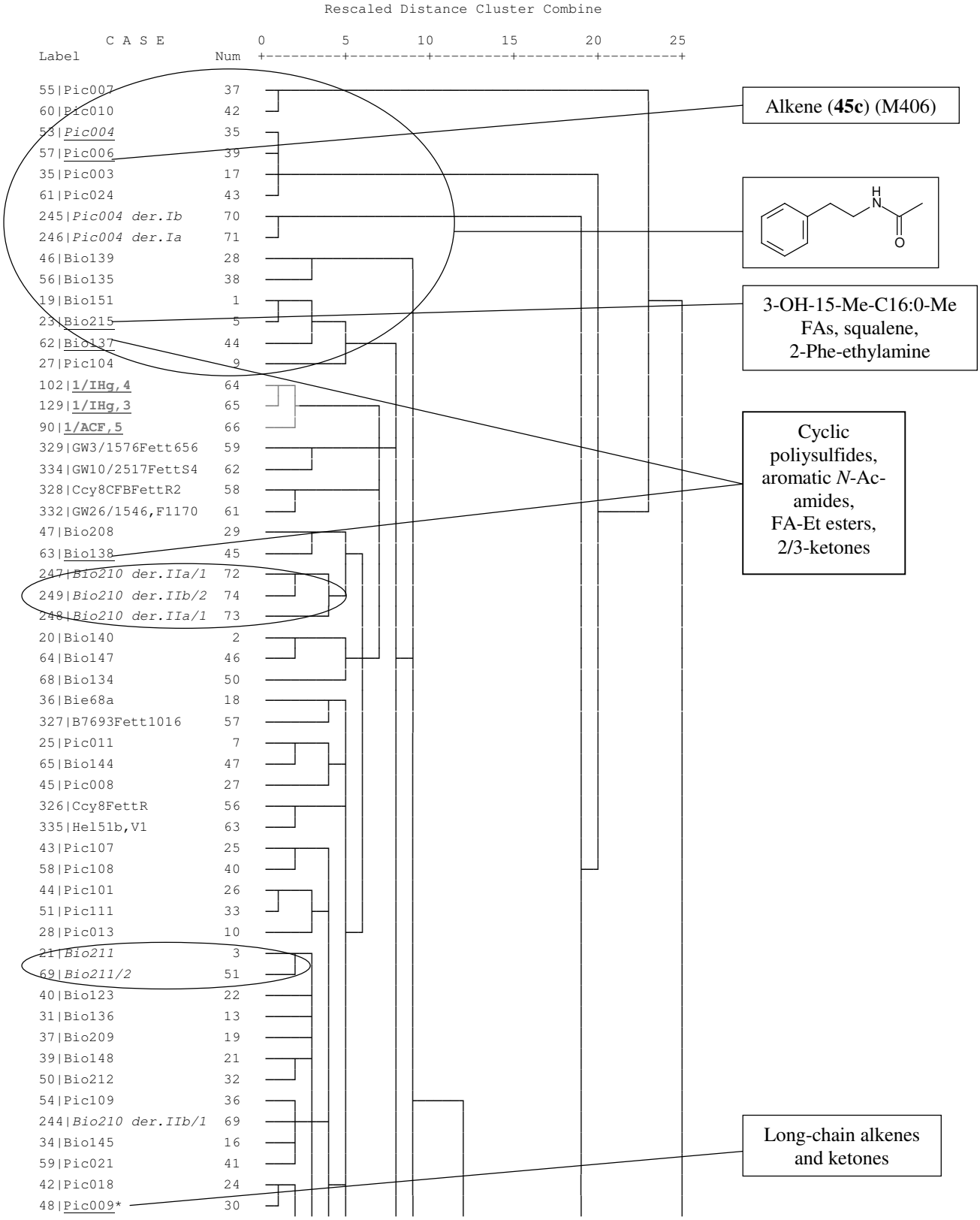
Dendogram 8.6 Subset STM3; PCA/HCA, 74% of the variance extracted, average linkage, quadratic Euclidean distance.



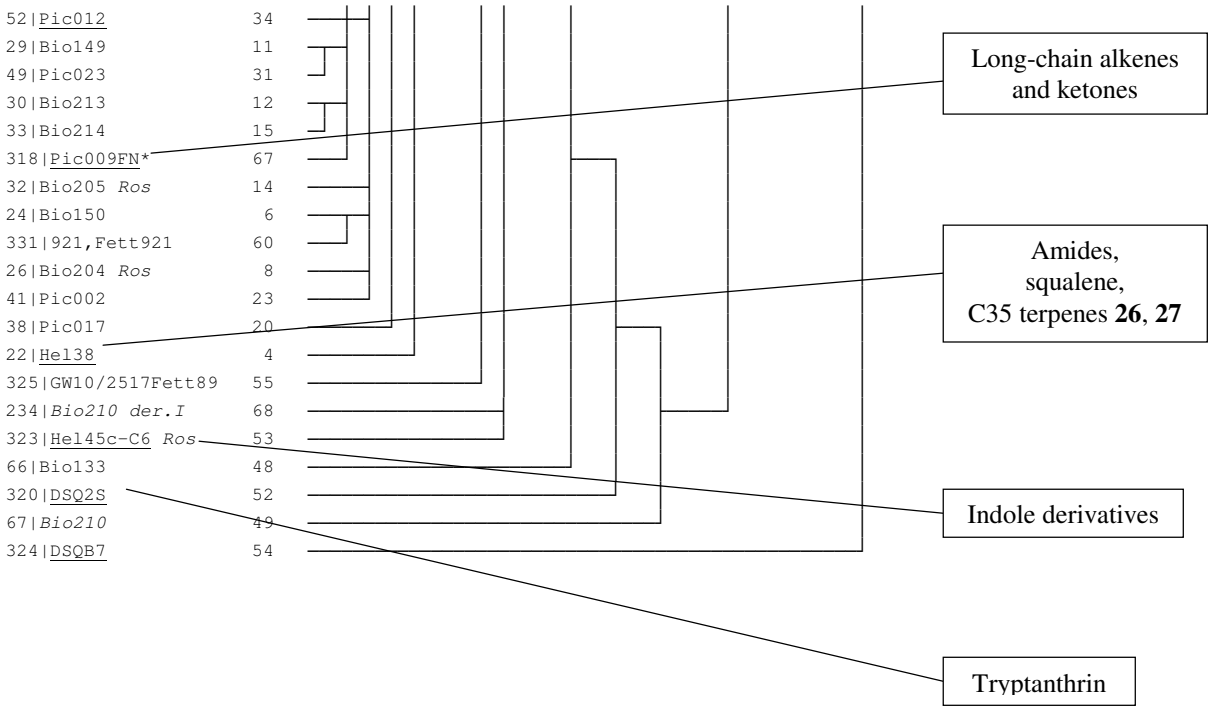
Dendogram 8.6 Continued.



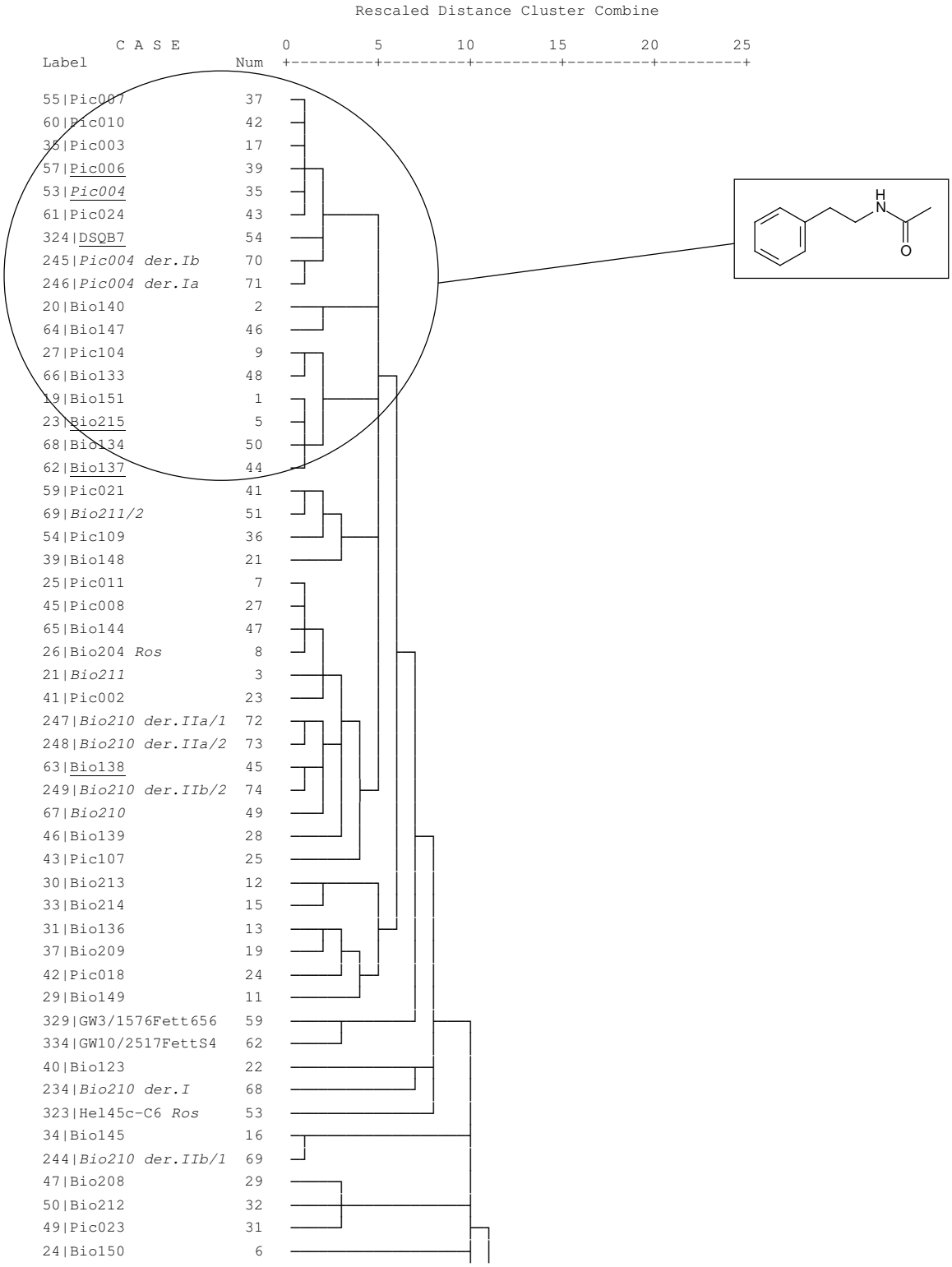
Dendogram 8.7 Subset STM4; HCA, average linkage, quadratic Euclidean distance.



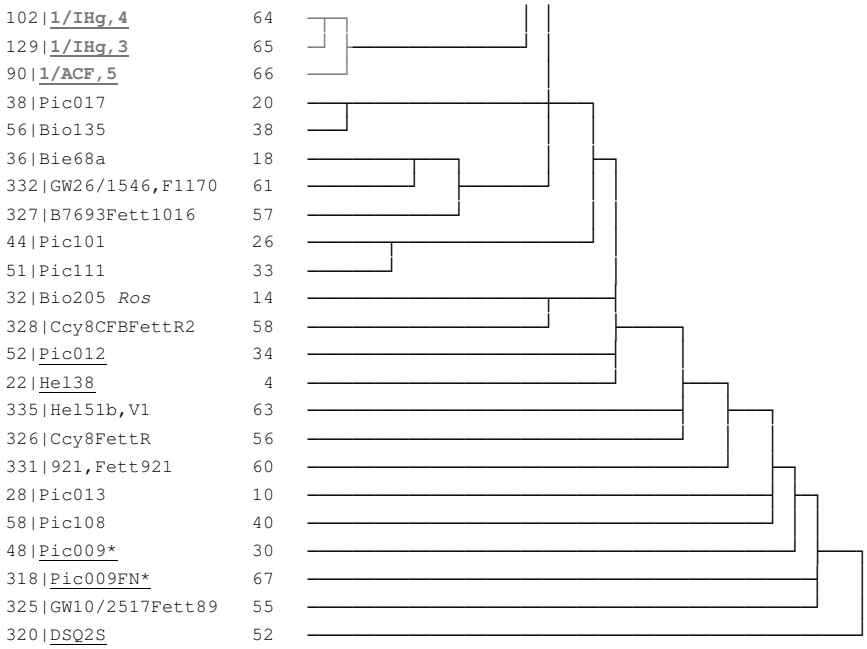
Dendogram 8.7 Continued.



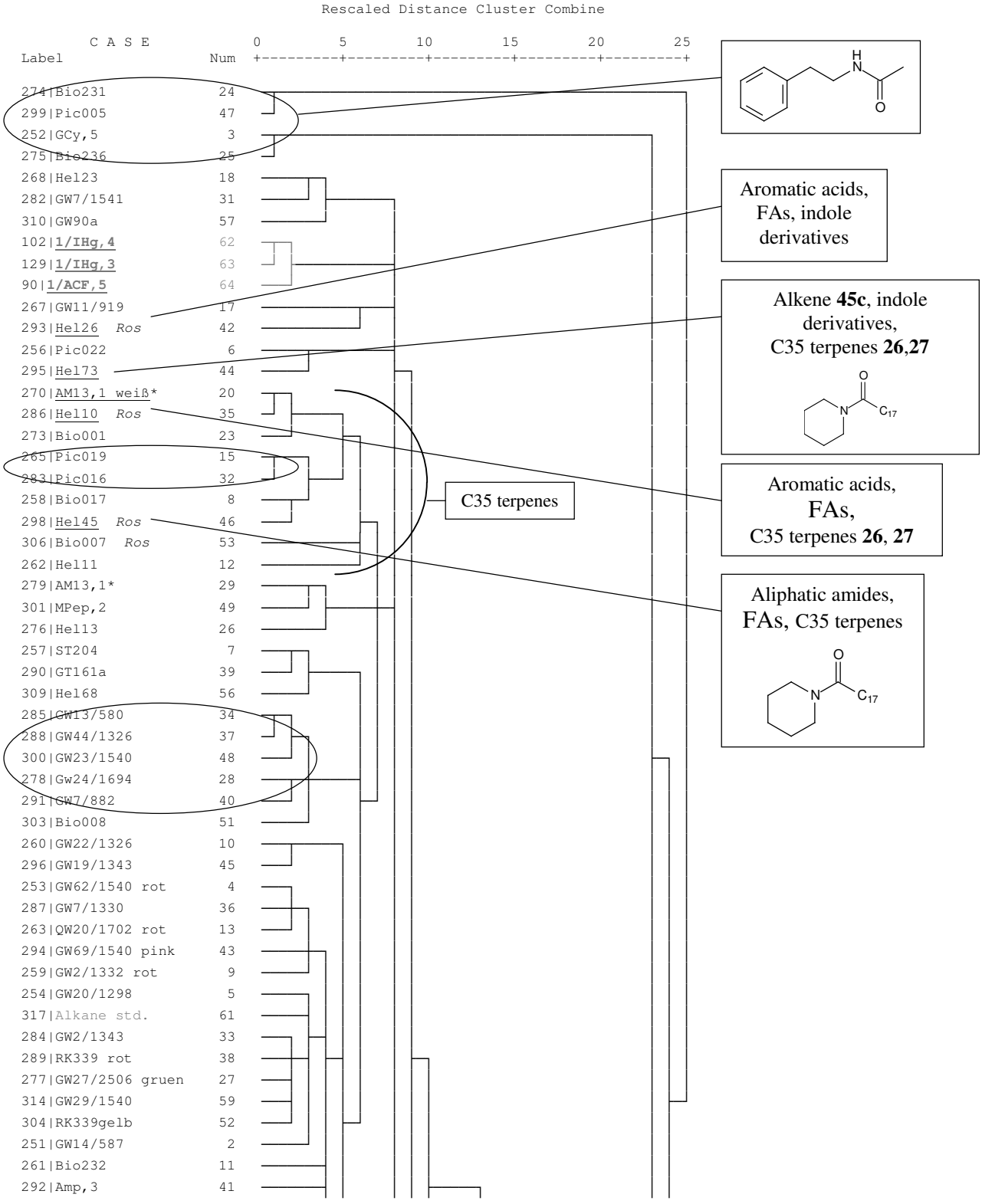
Dendogram 8.8 Subset STM4; PCA/HCA, 75% of the variance extracted, average linkage, quadratic Euclidean distance.



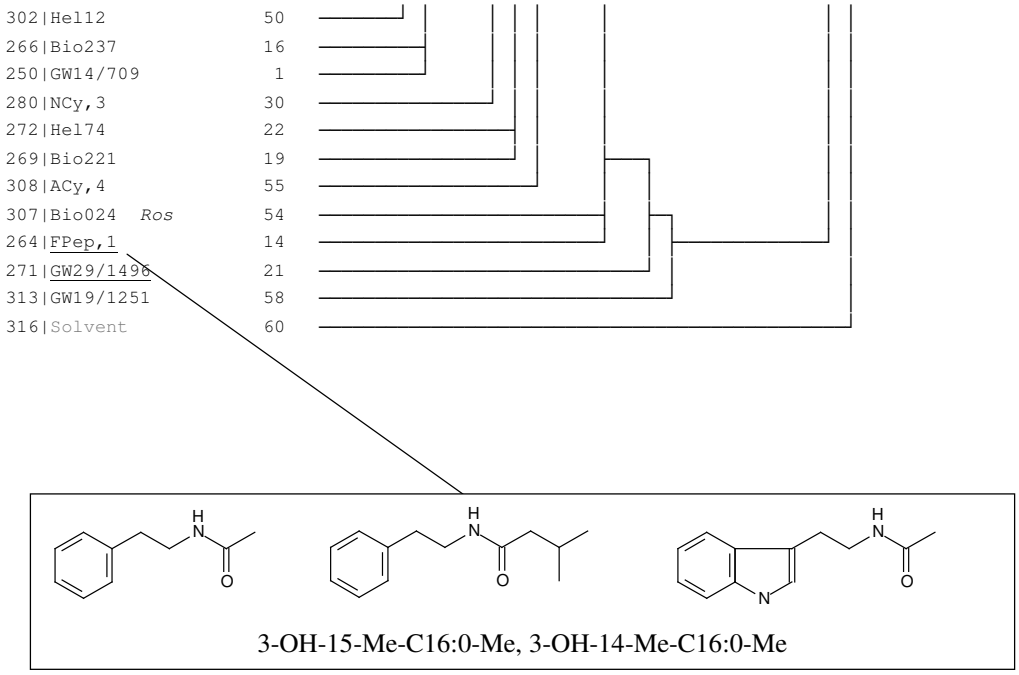
Dendogram 8.8 Continued.



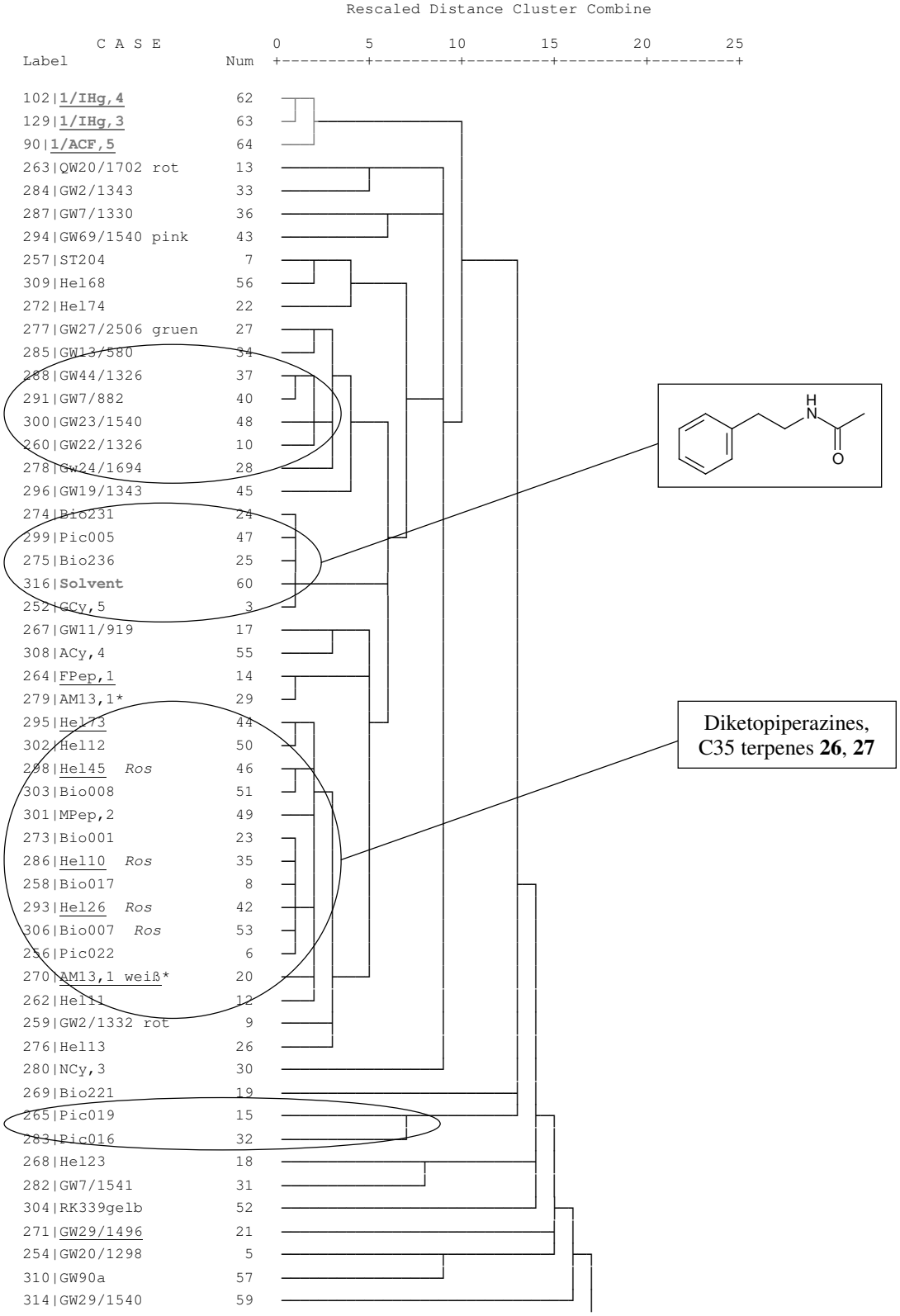
Dendogram 8.9 Subset STM5; HCA, average linkage, quadratic Euclidean distance.



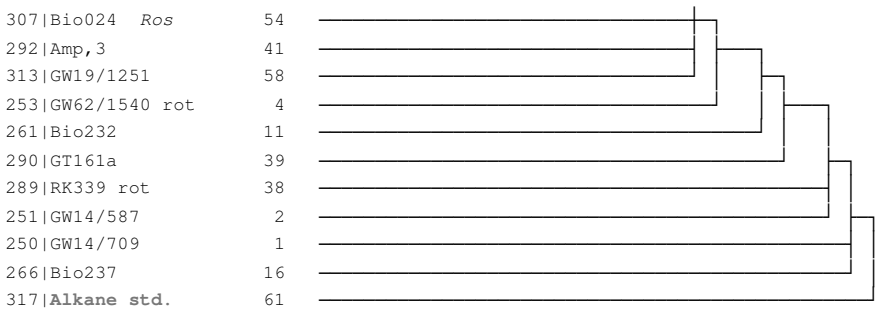
Dendogram 8.9 Continued.



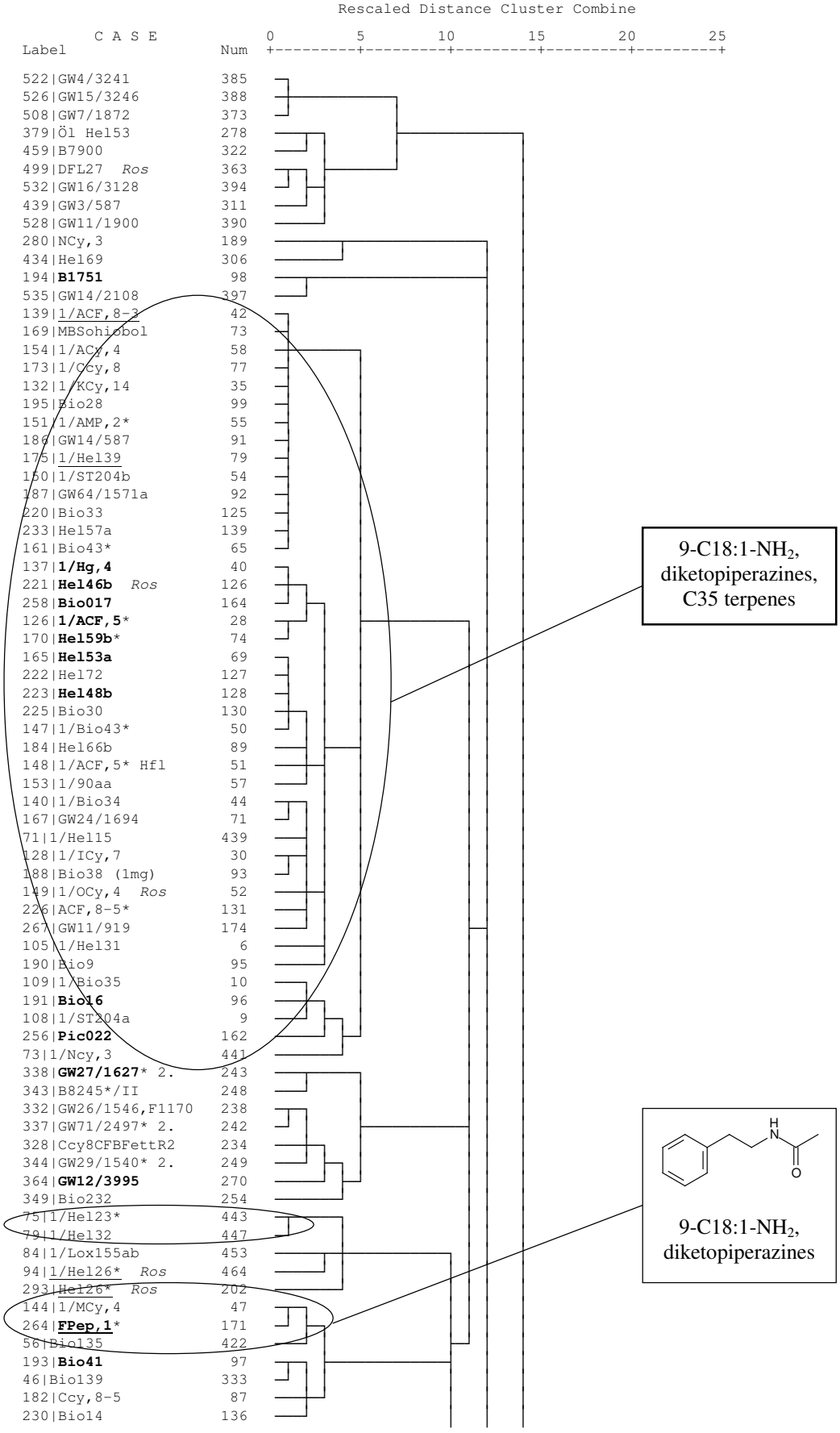
Dendogram 8.10 Subset STM5; PCA/HCA, 76% of the variance extracted, average linkage, quadratic Euclidean distance.



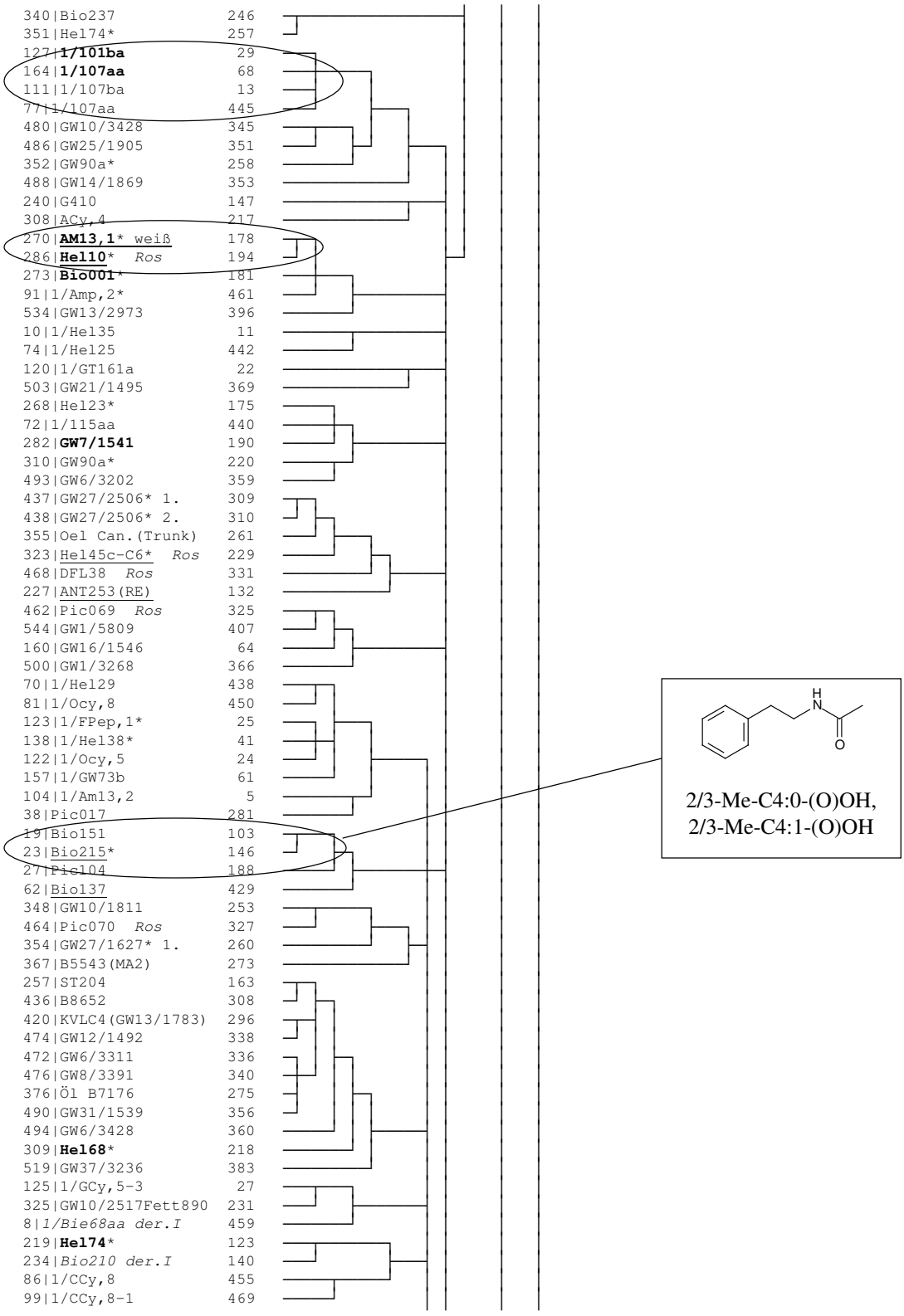
Dendogram 8.10 Continued.



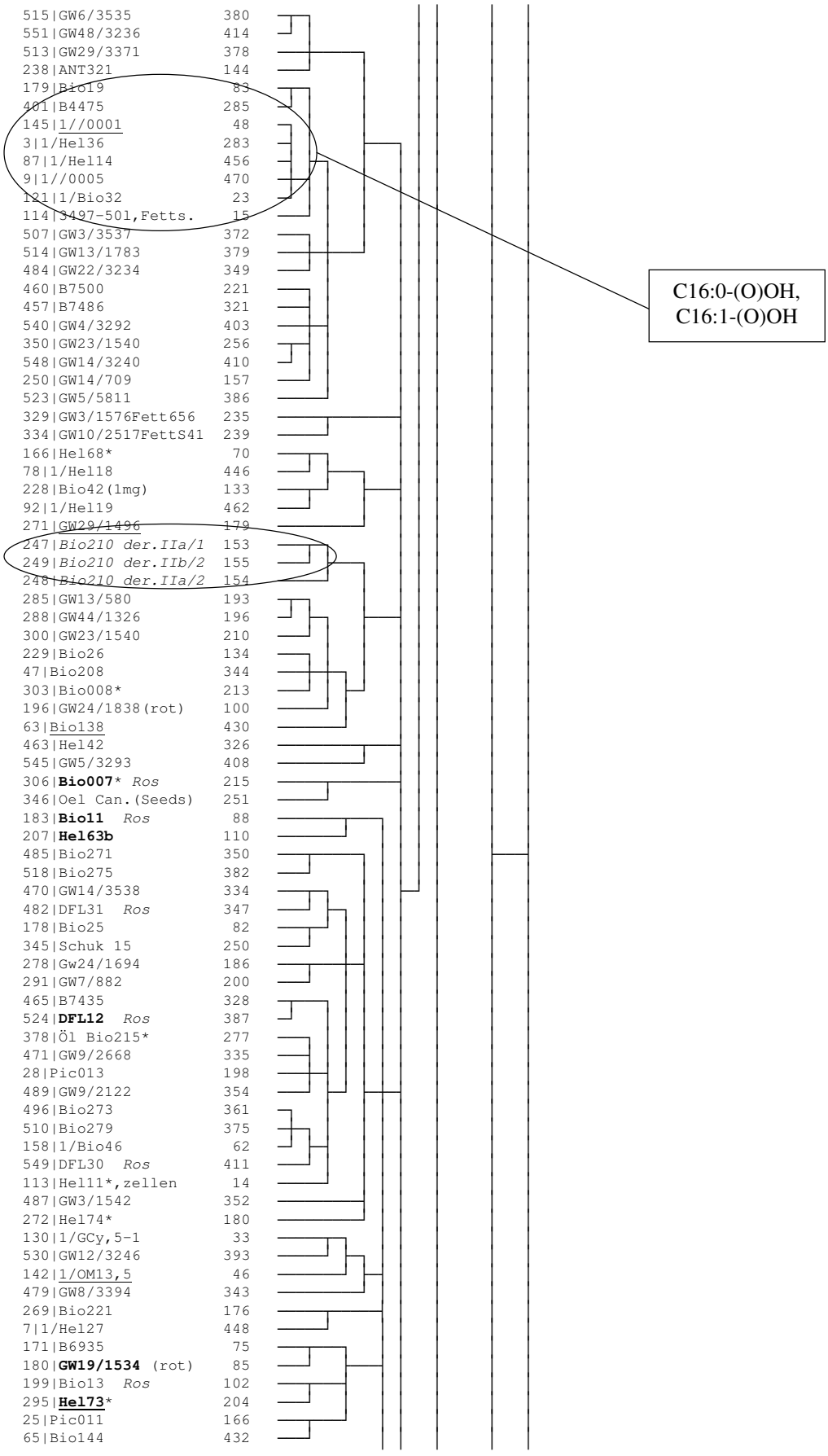
Dendogram 8.11 The entire data set of derivatized lipophilic extracts; HCA, average linkage, quadratic Euclidean distance; samples in bold contain the C35 terpenes **26** and **27**.



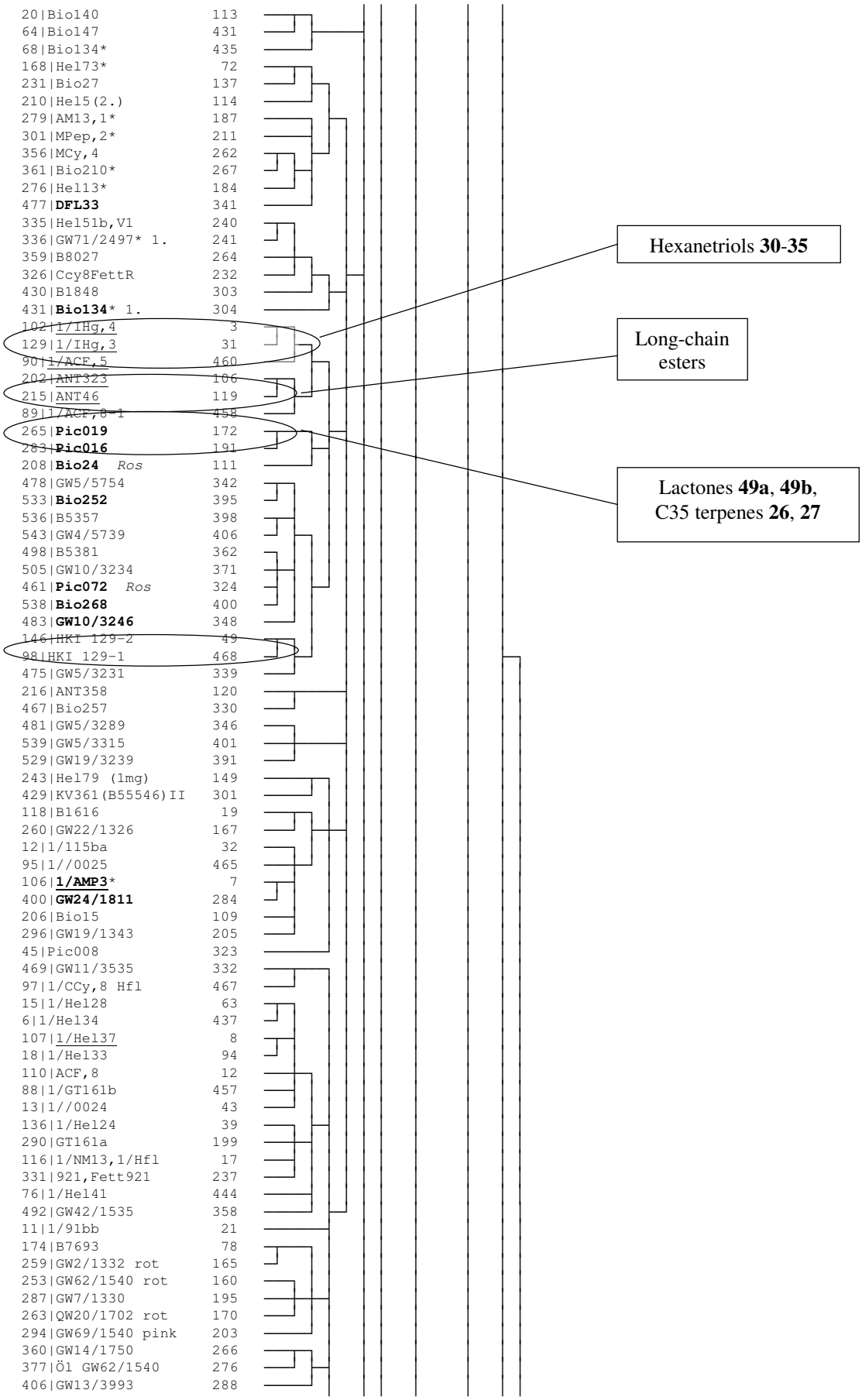
Dendogram 8.11 Continued.



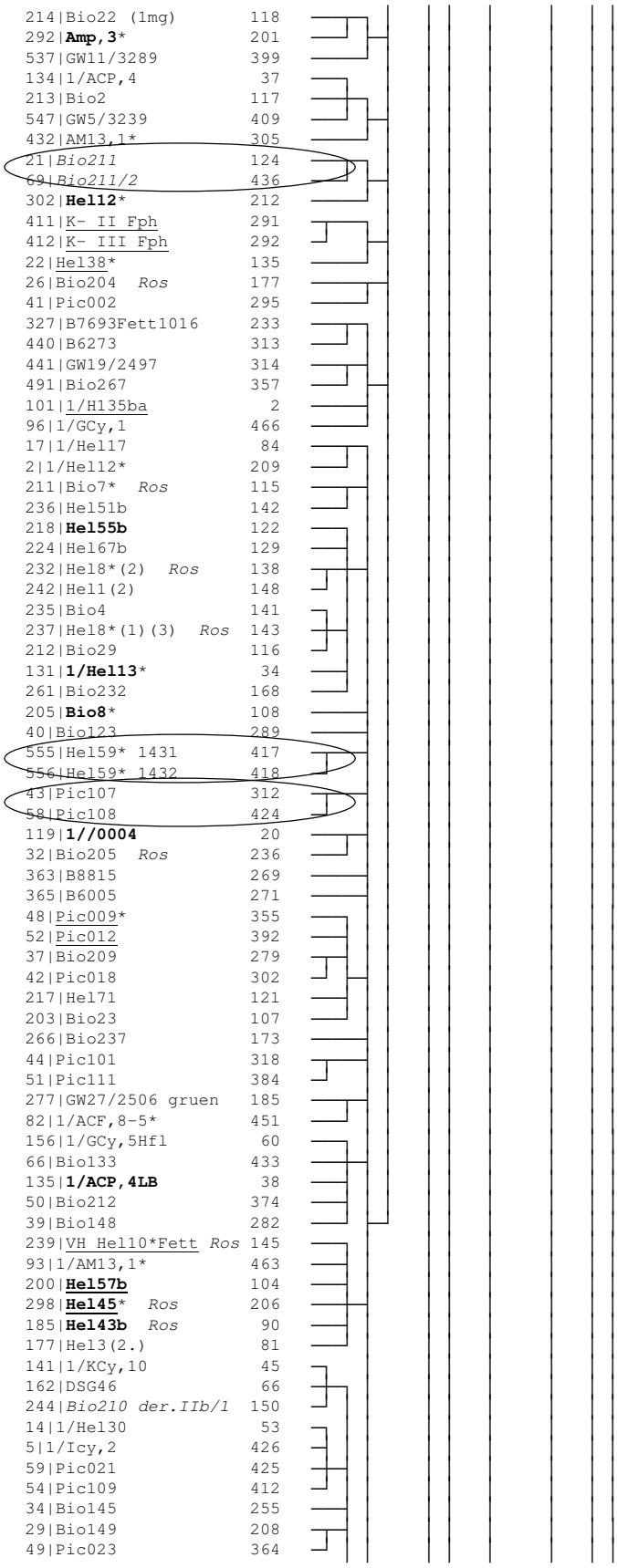
Dendogram 8.11 Continued.



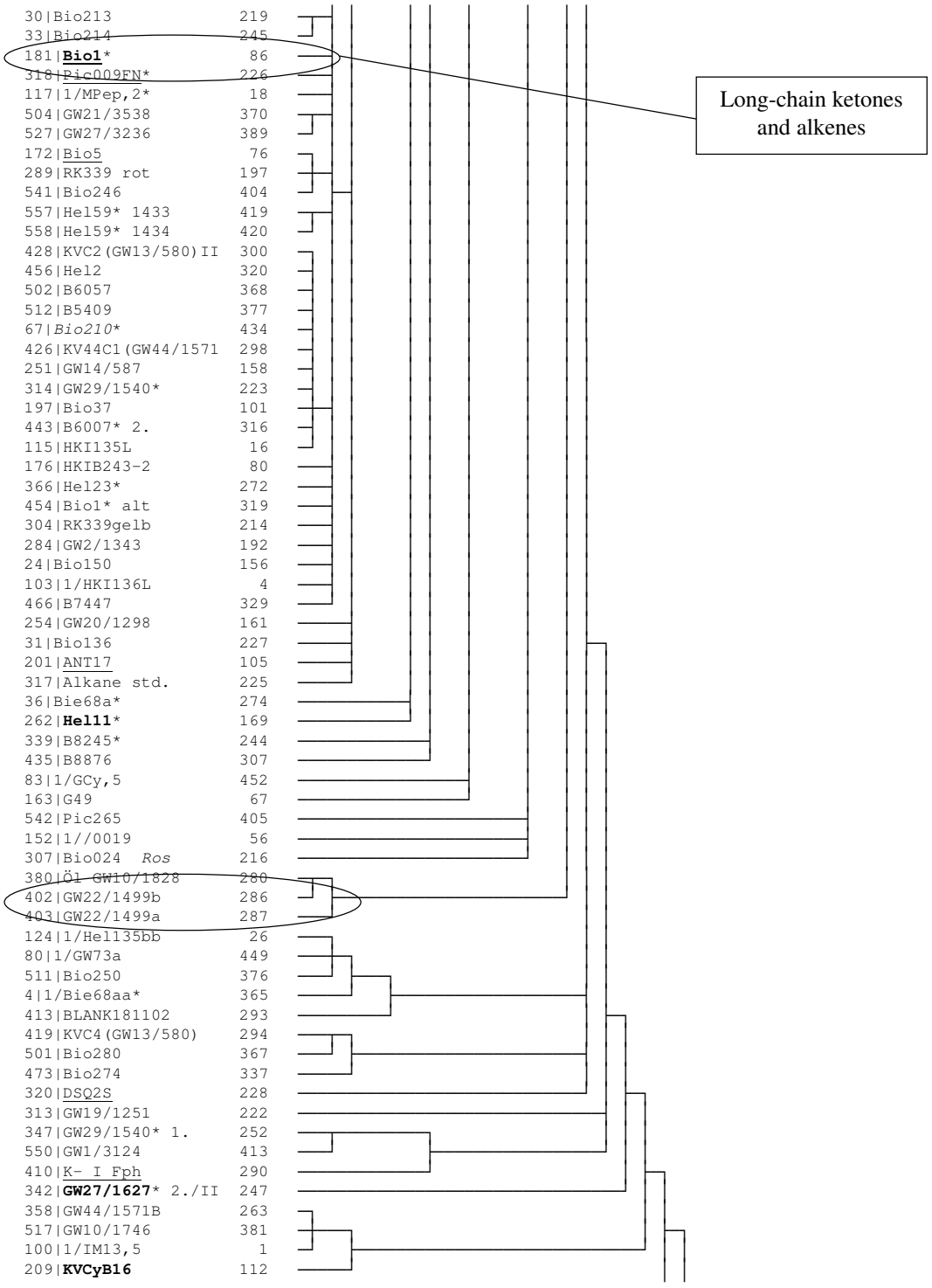
Dendogram 8.11 Continued.



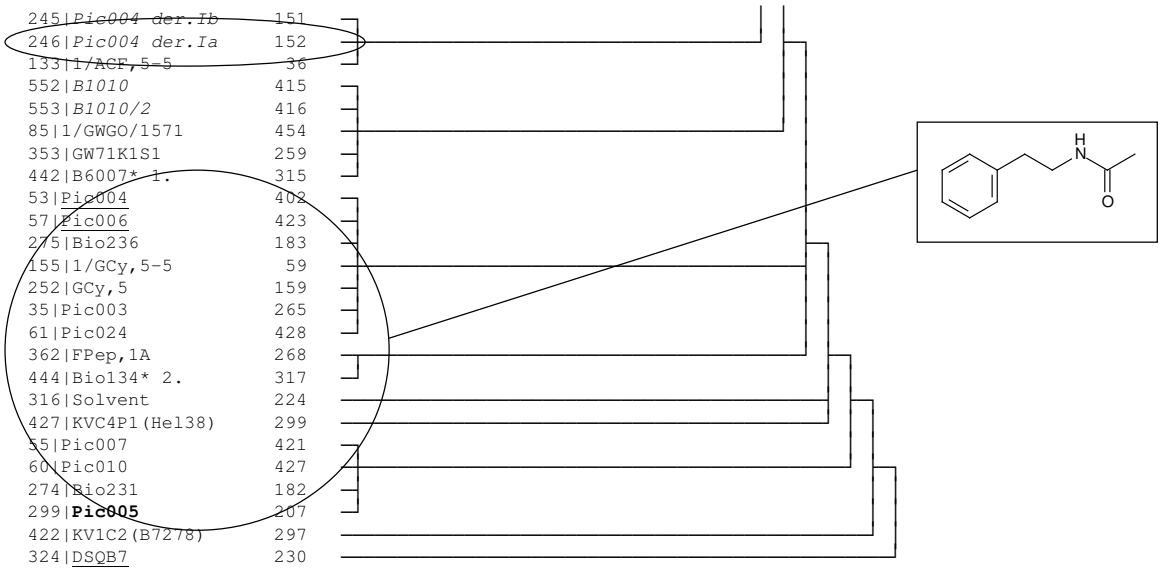
Dendogram 8.11 Continued.



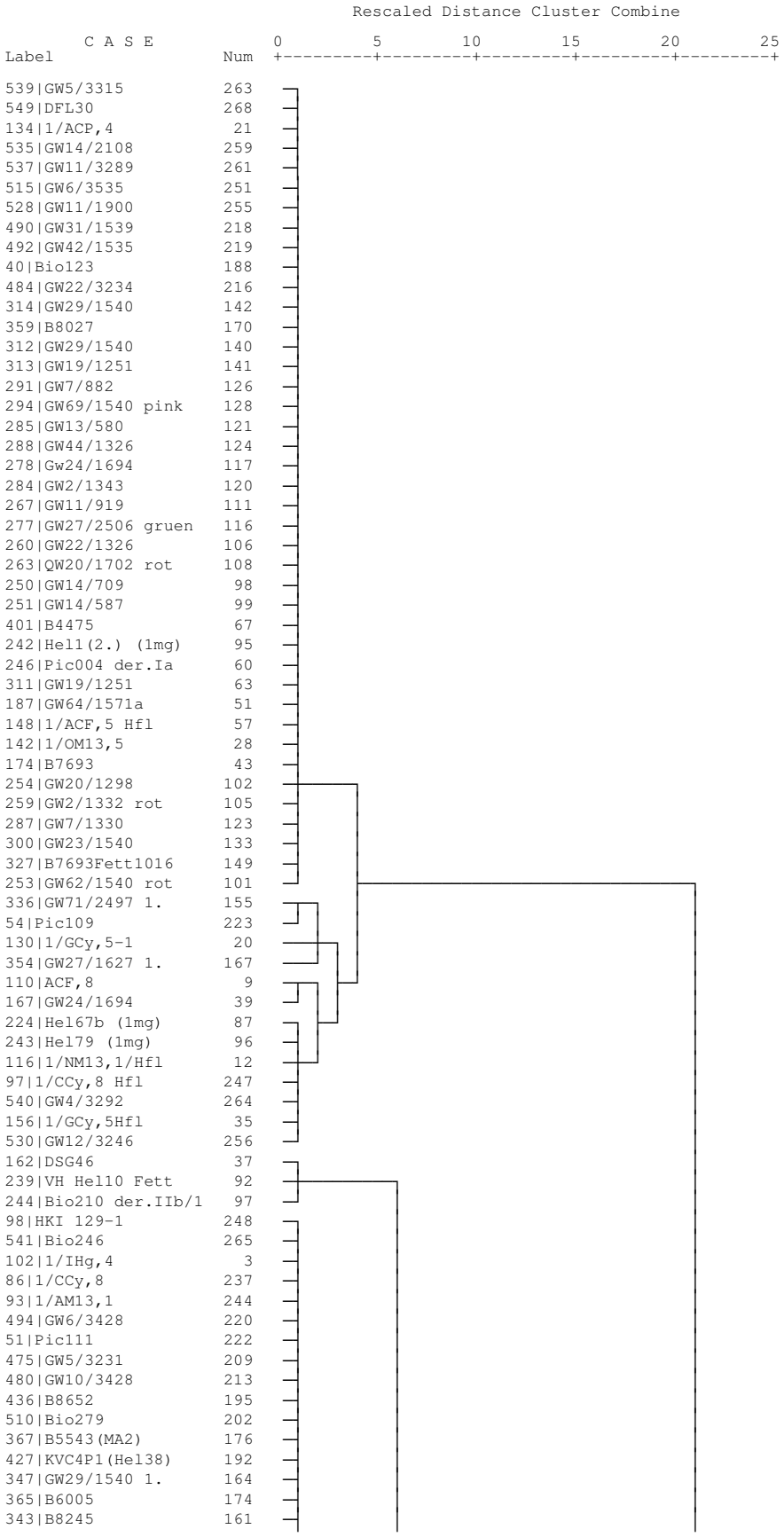
Dendogram 8.11 Continued.



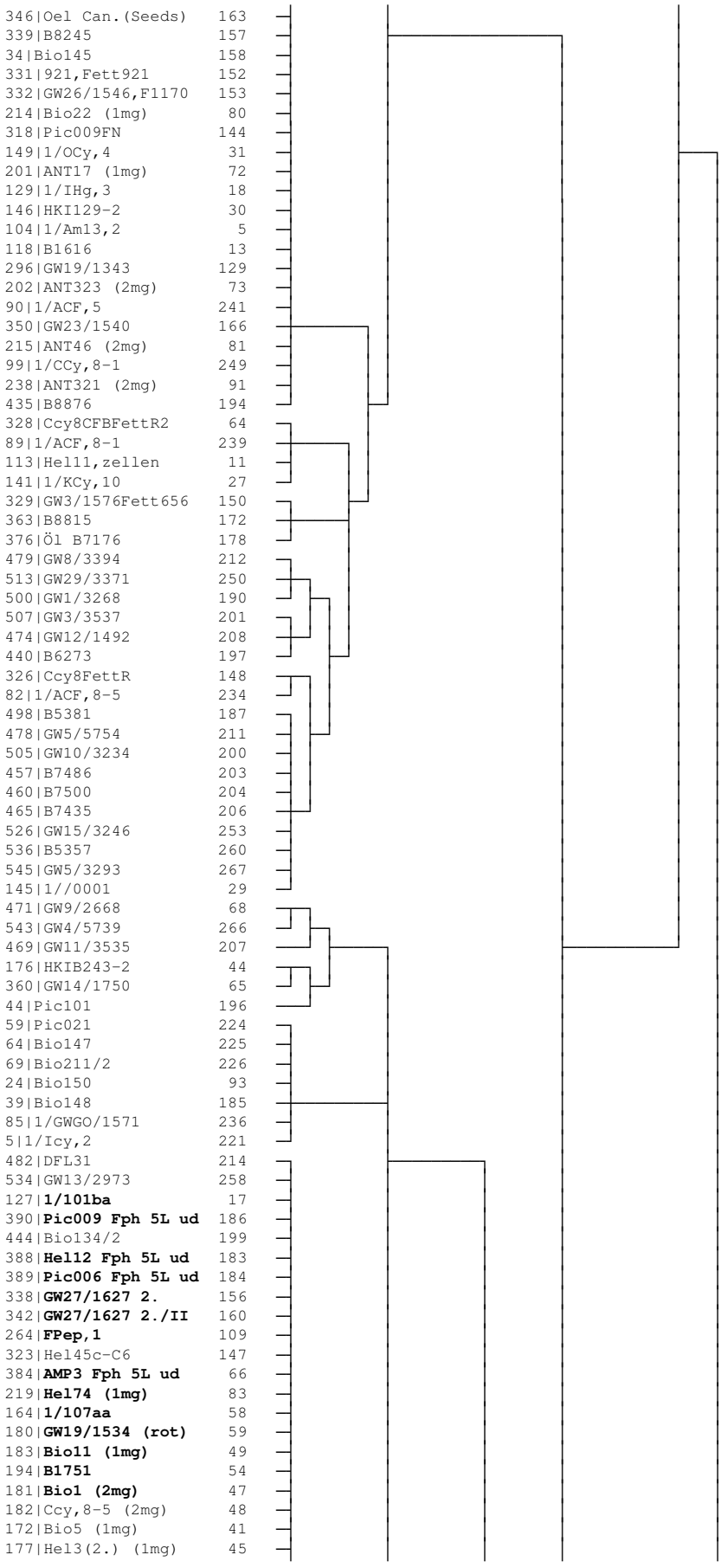
Dendogram 8.11 Continued.



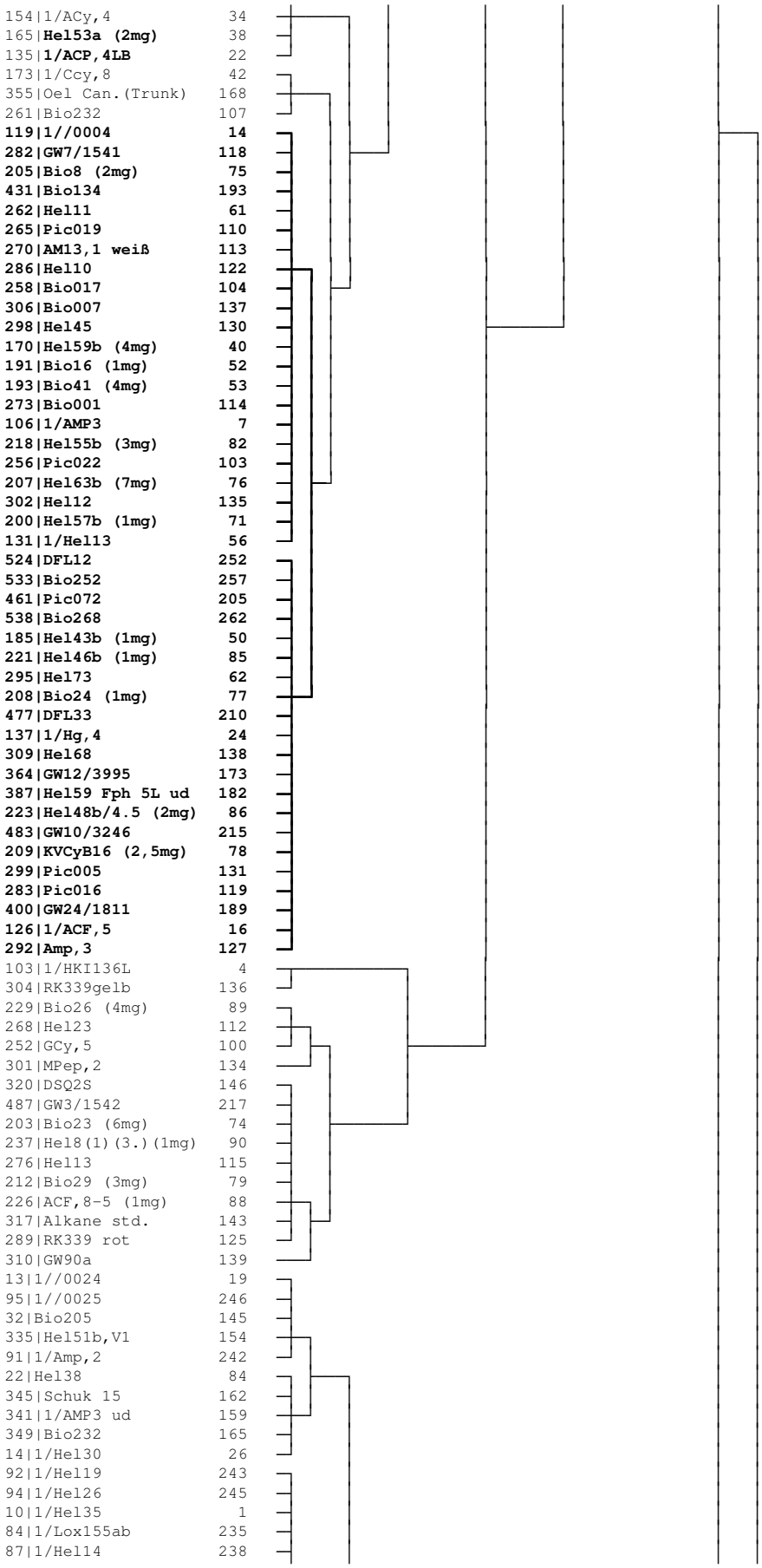
Dendogram 8.12 HCA of GC data from the time interval 84-89 min with all samples, average linkage, quadratic Euclidean distance; samples in bold contain the C35 terpenes **26** and **27**.



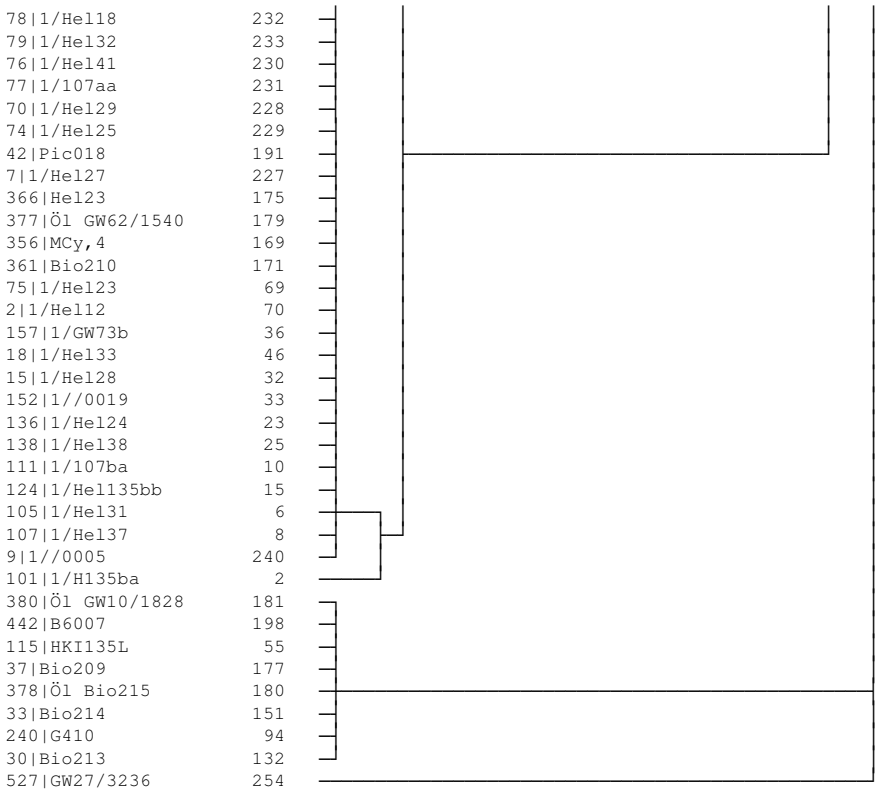
Dendogram 8.12 Continued.



Dendogram 8.12 Continued.



Dendogram 8.12 Continued.



DANKSAGUNG

Herrn Prof. Stefan Schulz danke ich für die Überlassung des Themas, die stete Diskussionsbereitschaft und die freundliche Zusammenarbeit.

Den Mitgliedern des Arbeitskreises Elena Barcari, Sonja Fietz-Razavian, Birte Flachsbarth, Jasmin Müller, Edyta Nagrodzka, Katja Stritzke, Verena Thiel, Nicole Thiesing, Selma Yildizhan, Rose-Maire Weiss, Christian Arsene, Gregor Brasse, Jeroen Dickschat, Jens Fuhlendorff, Stephan Goller, Christofer Häberlein, Nico Klewer, Karsten Krückert, Markus Müller, Andreas Schmidt, Paul Sobik und Robert Wegener danke ich für die Hilfsbereitschaft und die angenehme und freundliche Arbeitsatmosphäre.

Für die Bereitstellung des marinen Probenmaterials bedanke ich mich bei Prof. Laatsch der Universität Göttingen und bei Prof. Lang der TU Braunschweig.

Den Mitarbeitern der NMR-, und Massenabteilung und dem technischen Personal danke ich für ihre Unterstützung.

LEBENSLAUF

Persönliches

Name:	Katalin Böröczky
Geboren:	3.8.1975
	Budapest, Ungarn
Familienstand:	ledig

Bildungsgang

9.1993 – 6.1998	Studium an der Naturwissenschaftlichen Fakultät der Eötvös Loránd Universität, Budapest; Fachrichtung Chemie
3.7.1998	Diplom *Festphasenextraktion von organischen Spurenkomponenten aus verschiedenen Matrizen* Betreuer: Dr. Kornél Torkos
7.1997 – 8.1997	Praktikum bei Richter Gedeon AG, Budapest
9.2000 – dato	Dissertation am Institut für Organische Chemie der Technischen Universität Braunschweig *Chromatographic analysis of large numbers of marine bacteria extracts and the venom of the spider Cupiennius salei* Betreuer: Prof. Dr. Schulz

Berufspraxis

9.1998 – 7.2000	Budapester Wasserwerke AG, Budapest Abteilung für Wasserqualität und Umweltschutz Wissenschaftliche Mitarbeiterin
1.2000 – 2.2000	Bundesinstitut für gesundheitlichen Verbraucherschutz und Veterinärmedizin, Berlin Wissenschaftliche Mitarbeiterin